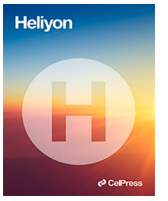
Heliyon 9 (2023) e21149

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Prediction of the interaction between *Calloselasma rhodostoma* venom-derived peptides and cancer-associated hub proteins: A computational study

Wisnu Ananta Kusuma [a,b,*], Aulia Fadli [a], Rizka Fatriani [b], Fajar Sofyantoro [c], Donan Satria Yudha [c], Kenny Lischer [d], Tri Rini Nuringtyas [c,e], Wahyu Aristyaning Putri [c], Yekti Asih Purwestri [c,e], Respati Tri Swasono [f]

[a] *Department of Computer Science, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, 16680, Indonesia*
[b] *Tropical Biopharmaca Research Center, IPB University, Bogor, 16128, Indonesia*
[c] *Faculty of Biology, Universitas Gadjah Mada, Yogyakarta, 55281, Indonesia*
[d] *Faculty of Engineering, University of Indonesia, Jakarta, 16424, Indonesia*
[e] *Research Center for Biotechnology, Universitas Gadjah Mada, Yogyakarta, 55281, Indonesia*
[f] *Department of Chemistry, Faculty of Mathematics and Natural Sciences, Universitas Gadjah Mada, Yogyakarta, 55281, Indonesia*

ARTICLE INFO

Keywords:
Bioinformatics
Biomedical
Cancer
Venom
Deep learning
Peptide

ABSTRACT

The use of peptide drugs to treat cancer is gaining popularity because of their efficacy, fewer side effects, and several advantages over other properties. Identifying the peptides that interact with cancer proteins is crucial in drug discovery. Several approaches related to predicting peptide-protein interactions have been conducted. However, problems arise due to the high costs of resources and time and the smaller number of studies. This study predicts peptide-protein interactions using Random Forest, XGBoost, and SAE-DNN. Feature extraction is also performed on proteins and peptides using intrinsic disorder, amino acid sequences, physicochemical properties, position-specific assessment matrices, amino acid composition, and dipeptide composition. Results show that all algorithms perform equally well in predicting interactions between peptides derived from venoms and target proteins associated with cancer. However, XGBoost produces the best results with accuracy, precision, and area under the receiver operating characteristic curve of 0.859, 0.663, and 0.697, respectively. The enrichment analysis revealed that peptides from the *Calloselasma rhodostoma* venom targeted several proteins (ESR1, GOPC, and BRD4) related to cancer.

1. Introduction

Patients with cancer are treated with three options, including surgery, radiotherapy, or pharmacotherapy, following their stage and tumor [1]. Recently, cancer treatment using targeted therapeutics in the form of peptide drugs is becoming an emerging trend that provides improved efficacy and reduced side effects [2]. Peptides have a crucial role in humans by interacting with various proteins and programming many cellular processes, such as cell death, and regulating gene expression [3].

* Corresponding author.Department of Computer Science, Faculty of Mathematics and Natural Sciences, IPB University, Bogor, 16680, Indonesia.
E-mail address: ananta@apps.ipb.ac.id (W.A. Kusuma).

https://doi.org/10.1016/j.heliyon.2023.e21149
Received 27 March 2023; Received in revised form 4 September 2023; Accepted 17 October 2023
Available online 26 October 2023

Furthermore, peptides have many advantages over small molecules and biologics, including simpler design, their ability to interact with unknown or unexplored targets, enhanced tissue penetration, and cheaper synthesis [1]. Research in peptides has become a good start for designing novel therapeutics due to their safety and tolerability in human bodies, and identifying peptide-protein interaction is crucial for this task [4]. However, finding new peptide-protein interactions often requires high costs and is very time-consuming [3]. Several methods have been developed to speed up peptide drug discovery.

Identifying peptide-protein interaction has two general approaches: sequence-based and structure-based. Sequence-based approaches use information in peptide and protein sequences to predict the interaction. For example, CGKronRLS [5] and NRLMF [6] calculate the similarity between sequences, and then use machine learning models to predict peptide-protein interaction. This approach generally needs known interaction between peptide and ligand and its similarity score as the input feature, which then created large-scale data and costs high computational power due to the enormous computational complexity of calculating similarity and the use of large-dimensional peptide and protein features [4]. Structure-based approaches, such as molecular docking, solved the problem by modeling the structure of peptides and proteins at the atom level and predicting their affinities. Several docking strategies to determine peptide-protein interaction can be divided into local and global docking methods. Most of the docking approaches need three-dimensional (3D) structure information from the peptide and protein to calculate the binding energies. The usage of 3D structures consumes large computational power, resources, and time [3].

The number of research and work regarding the identification of peptide-protein interaction using in silico approaches, such as machine learning or deep learning algorithm, remains relatively low although peptide drugs are becoming an emerging trend and the number of approved therapeutics using peptides has been increasing over the last decades. While the utilization of machine learning and deep learning techniques for peptide and protein interaction prediction is limited, these approaches have found significant application in the field of compound-protein interactions, specifically drug-target interactions (DTI). Several studies have successfully employed ensemble methods or deep learning methods to predict DTI. For instance, Ramadhanti et al. [7] utilized Random Forest to predict the compounds that interact with proteins in the SARS-COV-2 virus, focusing on COVID-19-related DTI. Sajadi et al. [8] introduced AutoDTI++, a deep unsupervised learning method that addresses the sparsity challenge in the interaction matrix and enables accurate prediction of drug-target interactions. Sulistiawan et al. [9] employed a stacked autoencoder (SAE) for pretraining the weights in a deep neural network (DNN), aiming to achieve higher prediction accuracy in DTI, particularly those related to COVID-19. Furthermore, Fadli et al. [10] conducted multilabel DTI prediction using the stack autoencoder-deep neural network (SAE-DNN) algorithm to address the limitations of binary classification models, which often overlook possible correlations between labels that could provide valuable information for enhancing the accuracy of DTI predictions.

Thus, our study aimed to utilize ensemble method and deep learning techniques to predict peptide-protein interactions using peptides derived from *Calloselasma rhodostoma* venom, with a particular focus on cancer-associated proteins. The methodologies employed included ensemble methods such as Random Forest (RF) and XGBoost, as well as a deep learning approach utilizing a deep neural network (DNN) that was enhanced with stacked autoencoders for pre-training (SAE-DNN). The goal was to leverage these techniques to accurately predict interactions between the venom-derived peptides and the target proteins associated with cancer. RF and XGBoost are bagging and boosting tree-based algorithms known for their good classification performance in bioinformatics and can handle high-dataset issues with the feature selection process even with a higher number of variables [11]. DNN is known for its ability to learn data representation and has achieved high performance in pattern recognition, computer vision, and bioinformatics [12,13]. We constructed two versions of the datasets: version 1 with a peptide sequence length of $\leq$50 and protein sequences length of $\leq$1500 and version 2 with a peptide sequence length of $\leq$19 (the average length of all peptide sequences) and protein sequences length of $\leq$1500. The feature profiles for peptides and proteins are constructed using intrinsic disorder, amino acid sequences, physicochemical properties, position-specific scoring matrix, amino acid composition, and dipeptide composition. We tested the peptide-protein interaction prediction model using the *Calloselasma rhodostoma* venom-derived peptides.

2. Materials and methods

2.1. Dataset

Three datasets were collected: the cancer protein dataset, the peptide-protein interaction dataset, and the peptide extracted from the *Calloselasma rhodostoma* venom, which will be used as test data to obtain new peptide-protein interaction. The cancer protein dataset is taken from The Cancer Genome Atlas website (https://www.cancer.gov/tcga accessed on October 3, 2022), resulting in a total of 571 proteins associated with cancer. The peptide data is collected from Protein Data Bank (PDB) [14] and DrugBank [15]. The peptide data in PDB is taken by first getting all the peptide and protein FASTA available, then using a protein-ligand interaction profiler (PLIP) to determine the interactions between peptide and protein. PLIP determines the interaction of peptide-protein by looking at seven types of interaction: hydrogen bonds, hydrophobic contact, pi-stacking, pi-caption interactions, salt bridges, water bridges, and halogen bonds [16]. Peptide-protein pairs are considered to have interaction if at least one of the interaction types is detected. Then, the generated interactions were filtered by selecting the cancer-associated proteins. We screened the drug-target interactions with peptide-type drugs and cancer proteins in DrugBank. A total of 452 positive interactions were obtained, including 404 from PDB and 48 from DrugBank.

Then, we generated the negative interactions to generate full datasets by randomly shuffling the peptides and proteins with unknown interactions. Five negative interactions were generated for each peptide by randomly sampling the non-interacting proteins. Overall, 2712 peptide-protein pairs were obtained in the training dataset.

We combined every unique venom peptide and cancer protein for the test data, which resulted in 23,925 peptide-protein pairs. The

Table 1
Encoded physicochemical feature.

Polarity	Positive/negative of hydropathy index	Amino acid	Encoded integer
Non-polar	Positive	Ala, Phe, Ile, Met, Leu, Pro, Val	1
Non-polar	Negative	Gly, Trp	2
Polar-uncharged	Positive	Cys	3
Polar-uncharged	Negative	Asn, Gln, Ser, Thr, Tyr	4
Negatively-charged	Negative	Asp, Glu	5
Positively-charged	Negative	Lys, His, Arg	6
Unknown	Unknown	Otherwise	7

test data consisted of 145 venom peptides. Venom was directly milked from *Calloselasma rhodostoma* snake. Freeze-dried venom was dissolved in buffer and digested using Trypsin Gold. The hydrolysate was measured for absorbance with a UV–Vis spectrophotometer at a wavelength of 585 nm. Then, trypsin-digested venom was subjected to high-resolution mass spectrometry. Hydrolysate was injected into the Acclaim® PepMap RSLC column and eluted using mobile phase A (water and 0.05 % formic acid) and B (water: acetonitrile at 20:80 and 0.1 % TFA). The MS and MS/MS were conducted using the NSI ionization method. Chromatograms of peptide fractions were obtained and analyzed using Xcalibur software. The dataset can be seen in Supplementary File 1.

2.2. Feature extraction

All peptide and protein sequences used have different lengths. We set the sequence to a certain length to make the same input feature size for all the peptides and proteins. We used two versions of data for the peptide data: peptide with a length of sequences of $\leq$50 [4] and peptide with a length of sequences of $\leq$19 (the average length of all peptide sequences). We set the length of protein sequences at$\geq$50 and $\leq$ 1500 for the protein data. Peptides with sequence lengths of $<$50 and 19 and proteins with $<$1500 amino acids were zero-padded to give the same input feature size. We then constructed the feature for both peptides and proteins using six feature profiles: amino acid sequences, physicochemical properties, intrinsic disorder, position-specific scoring matrix (PSSM), amino acid composition (AAC), and dipeptide composition (DPC).

We encoded each residue using an integer between 1 and 21 (the number of amino acid types) for the amino acid sequences (AAS) feature. We encoded residue in sequence using an integer between 1 and 7, which is based on the combination of polarity and the hydropathy index, for physicochemical property representation. This index measures the energy of the transfer of an amino acid side chain from a hydrophobic solvent to water [17]. Table 1 shows the encoded rules.

Intrinsic disorder (ID) features represent the tendencies of disordered amino acid pairs to form contacts with other objects such as protein or RNA. Its value ranges from 0 (complete order) to 1 (complete disorder). Three forms of intrinsic disorder scores are used: long disorder score (considering the long disorder regions), short disorder score (considering the short stretches of disorder), and ANCHOR score (the probability of specific amino acids to be a part of the disorder region) [18]. The use of intrinsic disorder for protein and peptide features can help increase the accuracy of peptide-protein interaction [4]. This feature is generated using IUpred2 [18].

PSSM feature is one the representation of protein sequences that can detect the homology between other protein sequences. This feature consists of an array S of size N $\times$ 20, with N as the length of the sequence and every position in Si, j as the probability of residue j in position i. PSSM profile is generated using psi-blast [19] with the number of iteration of 3 and e-value of 0.001 which have been proved to work well for generating effective feature profiles from the protein sequences [20]. The profile is then transformed using auto cross-covariance (ACC) after getting PSSM the profile of each peptide and protein. ACC can measure the correlation between any two properties and transform the PSSM profile into the same matrix. ACC is the combination of auto covariance between the same residue and cross-covariance (CC) between two different residues [21]. ACC transformation in PSSM profiles resulted in 400 features of peptide and protein. AAC and dipeptide composition represent the percentage of single amino acids and dipeptides (2 combinations of amino acids, such as AA, AR, and AN) in the sequences [22]. AAC converts the sequences into 20 features, while DPC converts them into 400 features.

2.3. Prediction model

The prediction model used three algorithms: deep neural network (DNN) with pre-training using stacked autoencoder (SAE), which will be referred to as SAE-DNN, RF, and XGBoost. SAE is used as pre-training to get the initial weight for DNN, which can help produce an optimal model compared to a model using random initial weights [23]. SAE training resulted in the weight and bias which will be used in the DNN training process. The activation functions in the output layer of DNN uses sigmoid because the data class is binary, while the other parameter will be tuned to find its optimal value.

RF is popular machine learning based on multiple decision trees (ensemble). RF achieves better prediction performance through the usage of bagging, randomly subsetting the variables, and a majority voting system [24]. RF can handle the high-dimensional dataset in bioinformatics by performing feature selection processes while building the prediction rules [25]. This study will tune several parameters of RF, namely: n_estimators (the number of trees built), max_samples (the number of variable samples used in the training process), and max_depth (the depth of individual trees).

XGBoost is a tree-based algorithm which is an optimization model that uses both the linear model and boosting tree model. XGBoost can improve the efficiency of the optimal results and achieve global optimal faster through the usage of the first derivative and second

Table 2
Scenarios details.

Scenario	Prediction model	Feature used
Model 1[a]	SAE-DNN	ID, PSSM
Model 2[a]	SAE-DNN	ID, PSSM, and Amino Acid Composition (AAC)
Model 3[a]	SAE-DNN	ID, PSSM, and Dipeptide Composition (DPC)
Model 4[a]	SAE-DNN	ID, PSSM, AAS, and PP
Model 5[a]	SAE-DNN	ID, PSSM, AAS, PP, and AAC
Model 6	SAE-DNN	ID, PSSM, AAS, and PP
Model 7	SAE-DNN	ID, PSSM, AAS, PP, and AAC
Model 8	SAE-DNN with RF feature selection	ID, PSSM, AAS, PP, and AAC
Model 9	SAE-DNN with XGBoost feature selection	ID, PSSM, AAS, PP, and AAC
Model 10	RF prediction model	ID, PSSM, AAS, PP, and AAC
Model 11	XGBoost prediction model	ID, PSSM, AAS, PP, and AAC
Model 12[b]	RF prediction model with class_weight settings	ID, PSSM, AAS, PP, and AAC
Model 13[b]	XGBoost prediction model with class_weight settings	ID, PSSM, AAS, PP, and AAC

[a] DrugBank dataset excluded.
[b] The weights are calculated using the inverse proportion of class frequencies.

Table 3
Hyperparameter search space.

Hyperparameter	Values
HL0 Node	100–2000
HLi Node	0.5 × (HL i-1)–0.75 × (HLi-1)
Hidden layer	1–6
Learning rate/eta[b]	0.01–0.1
Dropout rate	0.2–0.7
n_estimators[a b]	100–500
colsample_bytree[b]/max_samples[a]	0.1–1
max_depth[a]	3–6

[a] for the RF model.
[b] for XGBoost model.

derivative of the loss function for second-order derivation [26]. XGBoost also performs a feature selection process while being trained through the use of a gain score in each split tree and has already achieved good performance in the bioinformatics area [21].

This study attempted several scenarios, with scenarios 1–7 searching an optimal dataset for the model while scenarios 8–13 compared SAE-DNN, RF, and XGBoost performance. Table 2 shows the details of scenarios, with the only DrugBank data used from models 6–13. We used the peptide length of $\leq$50 and then compared the best model with the performance of the peptide length of $\leq$19 datasets to compare the performance of each scenario.

Hyperparameter tuning was also conducted using Bayesian Optimization to produce the optimal model. Table 3 shows the hyperparameter search space.

2.4. Model evaluation

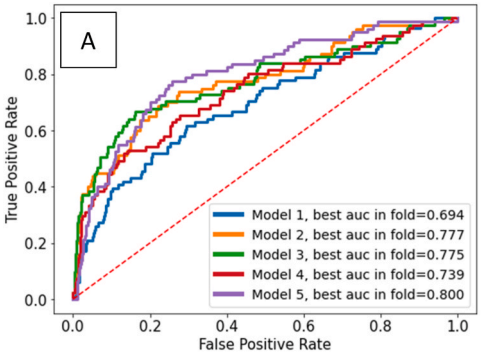
The model is trained using StratifiedKFold Cross Validation to balance the number of positive and negative interactions, with the number of folds, K = 5. The evaluation metrics include accuracy, recall, precision, Fscore, receiver operating characteristic curve (ROC), and precision-recall curve (PRC). The accuracy value measures how well the test data predicts. Precision measures the percentage of positive predictions against a positive class. Recall measures the accuracy of the positive prediction of the model. Fscore measures the performance of the minority class. The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 - FPR) [9]. The PRC curve shows the comparison between precision and recall which aimed to measure how well the model predicted the class in imbalanced settings [27].

2.5. Enrichment analysis

Enrichr (https://maayanlab.cloud/Enrichr/) [28] is a database used to search for protein pathway enrichment. Pathway enrichment analysis consists of identifying the KEGG Pathway and Gene Ontology (Biological Processes, Molecular Functions, and Cellular Components). The online tool SRPlot (http://www.bioinformatics.com.cn/). Cytoscape v3.9.1 (https://cytoscape.org/) [29] was used to build protein-enrichment and protein-peptide-enrichment network topologies to visualize predicted results of pathway enrichment.

Table 4
All scenario comparison results with peptide length of ≤50 dataset.

Metrics		Accuracy	Recall	Precision	ROC-AUC	Fscore
Model	**1**	0.825 ± 0.005	0.136 ± 0.046	0.429 ± 0.040	0.647 ± 0.030	0.202 ± 0.053
	2	0.831 ± 0.024	0.341 ± 0.062	0.509 ± 0.094	0.723 ± 0.036	0.402 ± 0.058
	3	0.832 ± 0.016	0.428 ± 0.047	0.502 ± 0.052	0.733 ± 0.031	0.460 ± 0.042
	4	0.825 ± 0.020	0.349 ± 0.062	0.494 ± 0.091	0.726 ± 0.013	0.397 ± 0.035
	5	0.822 ± 0.016	0.495 ± 0.014	0.485 ± 0.089	0.761 ± 0.024	0.486 ± 0.047
	6	0.804 ± 0.003	0.433 ± 0.079	0.428 ± 0.011	0.654 ± 0.023	0.428 ± 0.043
	7	0.818 ± 0.020	0.358 ± 0.118	0.480 ± 0.094	0.736 ± 0.025	0.393 ± 0.084
	8	0.828 ± 0.012	0.427 ± 0.072	0.499 ± 0.044	0.743 ± 0.013	0.457 ± 0.051
	9	0.830 ± 0.017	0.413 ± 0.100	0.518 ± 0.055	0.753 ± 0.017	0.448 ± 0.050
	10	0.842 ± 0.007	0.148 ± 0.012	0.709 ± 0.133	0.688 ± 0.039	0.243 ± 0.021
	11	0.859 ± 0.006	0.369 ± 0.040	0.663 ± 0.055	0.697 ± 0.020	0.466 ± 0.025
	12	0.800 ± 0.014	0.433 ± 0.067	0.419 ± 0.037	0.685 ± 0.042	0.455 ± 0.032
	13	0.843 ± 0.009	0.400 ± 0.023	0.560 ± 0.038	0.690 ± 0.029	0.466 ± 0.023



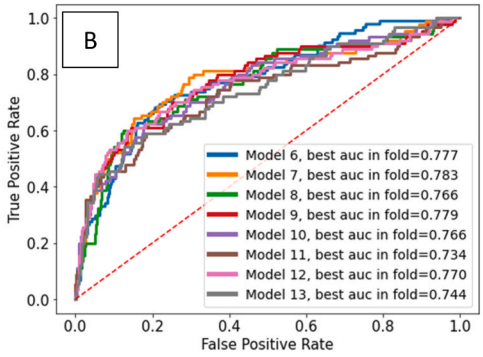
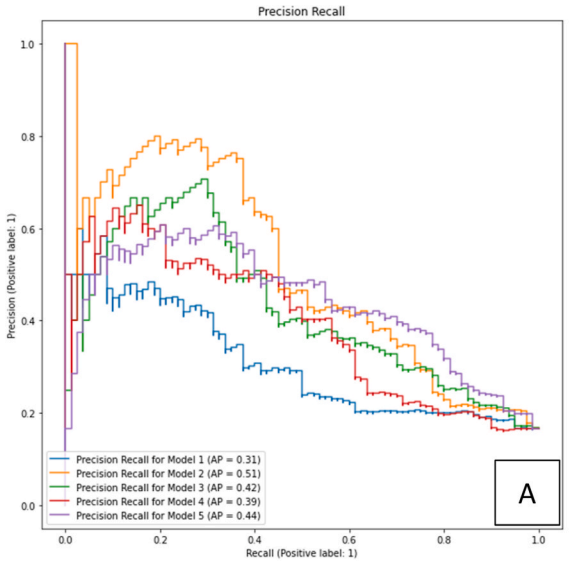


Fig. 1a. ROC curve for model 1–5.
Fig. 1b. ROC curve for model 6-13.



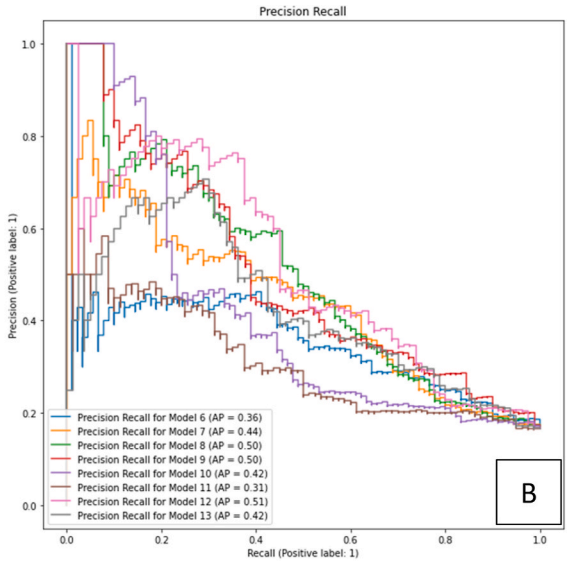


Fig. 2a. Precision-Recall (PR) curve for model 1–5.
Fig. 2b. Precision-Recall (PR) curve for model 6-13.

Table 5
All scenarios' confusion matrix comparison results.

Metrics	Model												
	1	2	3	4	5	6	7	8	9	10	11	12	13
True Positive (TP)	11	32	29	31	40	37	17	41	31	13	32	34	33
True Negative (TN)	392	387	374	374	360	390	428	408	418	437	427	390	413
False Negative (FN)	69	48	51	49	40	53	73	49	59	77	58	56	57
False Positive (FP)	12	17	30	30	44	48	10	30	20	1	11	48	25

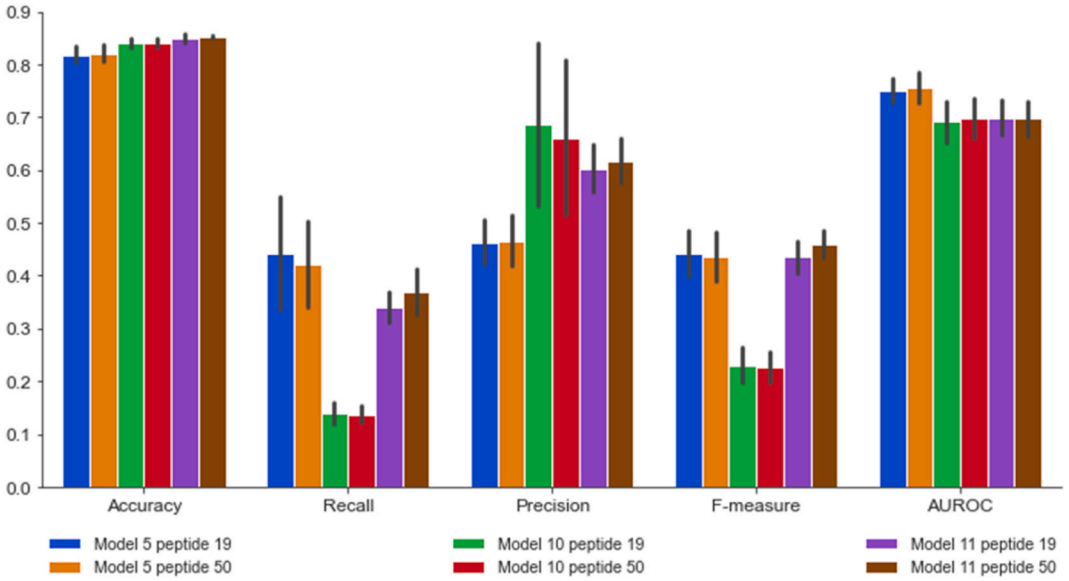


Fig. 3. Comparison results for dataset versions 1 and 2 using models 5, 10, and 11.

3. Results and discussion

3.1. Prediction results of dataset version 1

Supplementary Table 4 shows all the prediction models trained using the optimal hyperparameters. Table 4 shows all scenario comparison results.

All models generally produced fairly good average accuracy with a score of ≥80 %. All models have a decent result in an average ROC-AUC of >0.6. However, all models produced a bad recall and Fscore, with a score of under 0.4 in both metrics. A low recall score indicated a high number of false negatives in the models which is a common case in imbalanced class problems.

The ROC curve is used for another measurement to evaluate the model. The ROC curve of each model can be seen in Fig. 1a and b. The ROC curve closer to the top-left corner indicates the following. Higher AUC means that the model can distinguish classes between positive and negative. Hence, model 5 has the best AUC in one of its folds of 0.8, while the other models produced almost similar scores of AUC with only model 1 scoring below 0.7. This means that all scenarios, except model 1, have at least a 73 % chance to differentiate the class between positive and negative.

An imbalanced setting is PRC is another metric used to measure model performance, which can be seen in Fig. 2a and b. PRC shows the trade-off between precision and recall, and a model is considered to perform better when moving closer to point (1,1) in the graph. Hence, no models tend to be close to the ideal point with some of the models having a high precision and low recall. A model with high precision but low recall means returning very few positive classes, but most of its predictions were correct.

The scenarios were further compared by measuring the confusion matrix on the last fold of the CV. Table 5 shows a comparison of the results.

Models 5 and 8 had the highest number of TP (40 and 41 consecutively), but they also produced many FP which is something that should be avoided in peptide-protein interactions. Model 10 produced the lowest number of positive predictions with only 14. Hence, we considered models 5, 10, and 11 to be the better model among others with model 11 being the best due to its high number of TP (32) and low FP (11) as well as its performance in all evaluation metrics in Table 4.

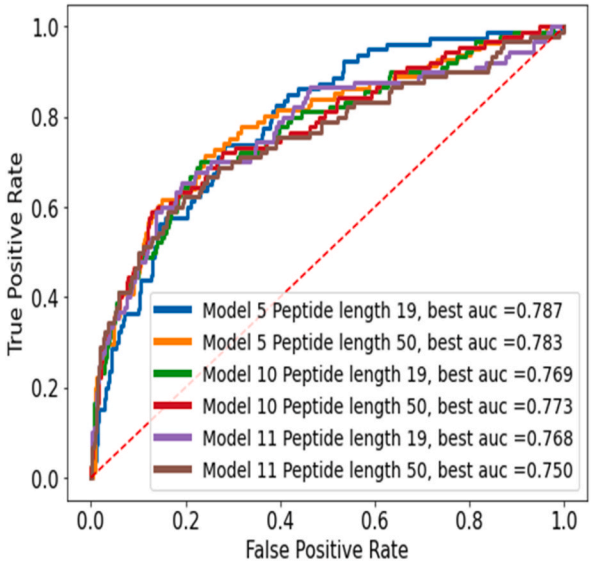


Fig. 4. ROC curve comparison for dataset versions 1 and 2 using models 5, 10, and 11.

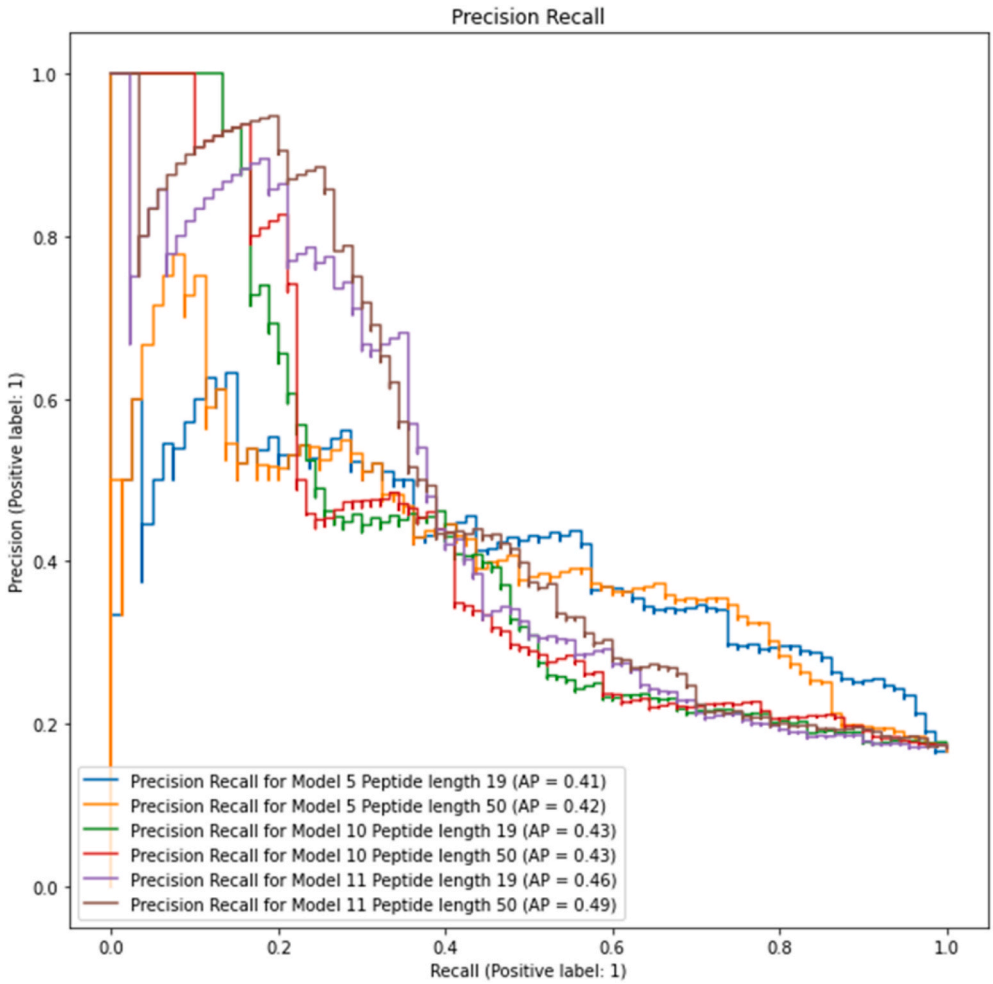


Fig. 5. Precision-Recall curve comparison for dataset versions 1 and 2 using models 5, 10, and 11.

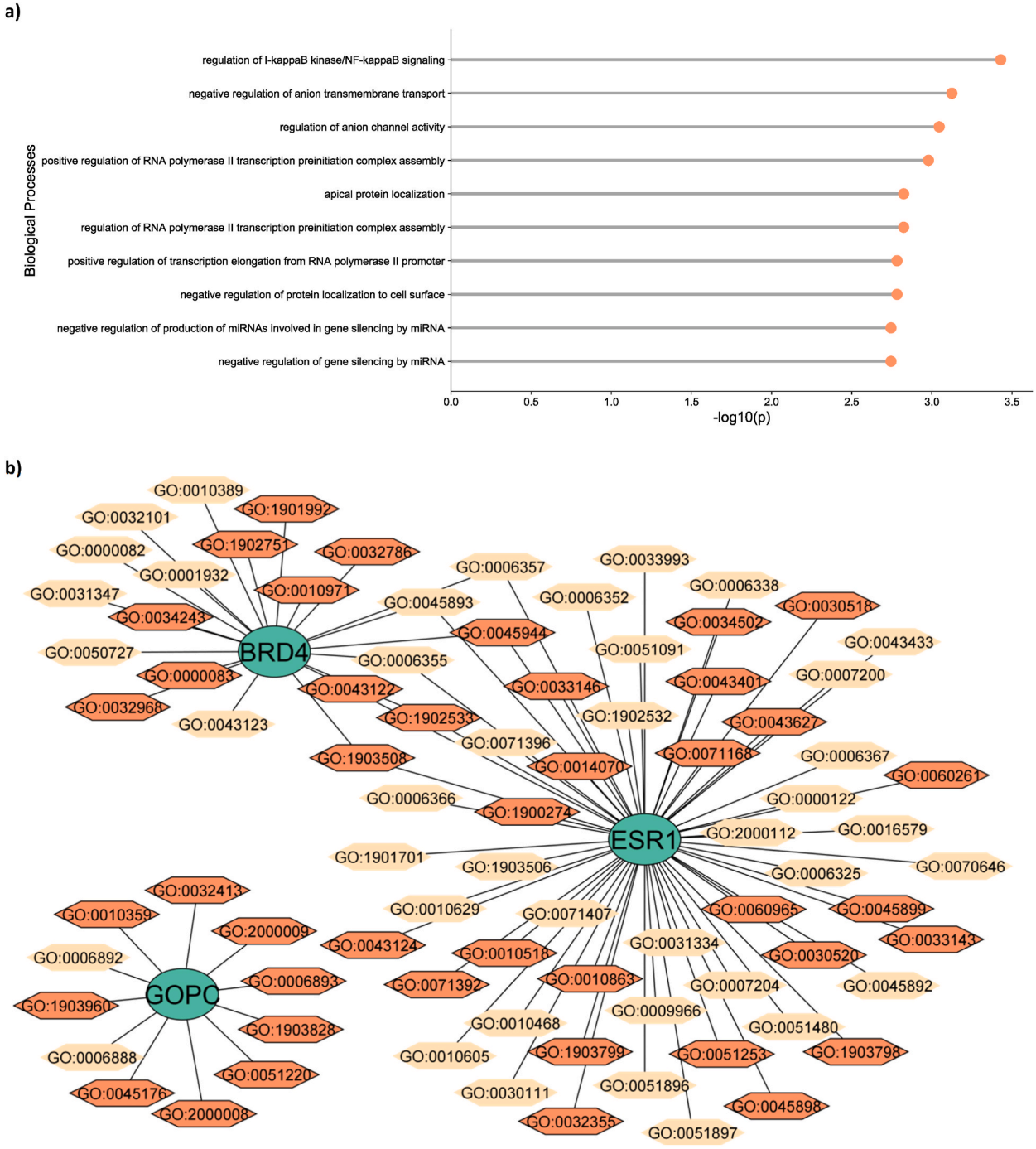


Fig. 6. GO Biological Processes: a. Top 10 terms with the lowest *p*-value; b. network of protein-GO Biological Processes, darker colors indicate the terms with a *p*-value of <0.01(enrichment terms explanation can be seen in Supplementary File 5). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3.2. Prediction comparison of dataset version 1 and version 2 using the three best model

Then, we compared the performance of models 5, 10, and 11 with the data version 2 which used the peptide sequence length of $\leq$19 (the average length of all peptide sequences). The comparison results can be seen in Fig. 3.

Overall, the performance of dataset versions 1 and 2 revealed no significant differences. Dataset version 2 has a slightly better performance in model 10 while dataset version 1 had a better performance in models 5 and 11. Fig. 4 shows the comparison for the

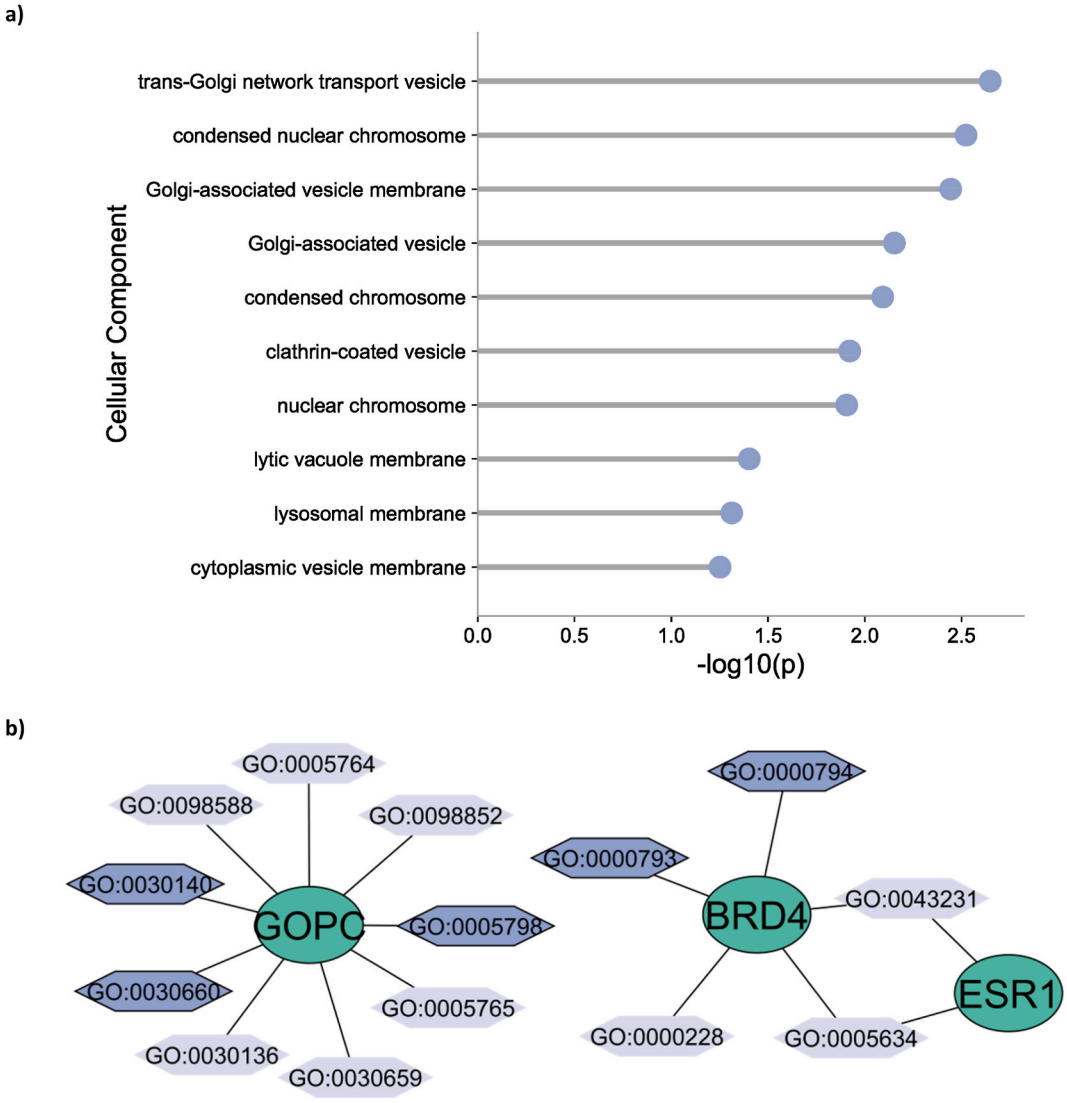


Fig. 7. GO Cellular Component: a. Top 10 terms with the lowest *p*-value; b. network of protein-GO Cellular Component, darker colors indicate the terms with a *p*-value of <0.01 (enrichment terms explanation can be seen in Supplementary File 5). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

ROC curve, while Fig. 5 shows the PR curve comparison of each dataset.

3.3. Venom peptide prediction results in dataset version 1

The venom peptide prediction is performed using models 5, 10, and 11. Prediction using model 5 resulted in 23,532 pairs of peptide-protein with all peptides predicted to have interaction with cancer proteins. Only protein AFDN had no interaction with venom peptide.

Prediction using model 10 produced 169 pairs of peptide-protein with 113 unique peptides predicted to have an interaction with only four proteins (Estrogen Receptor 1 [ESR1], Golgi-associated PDZ and coiled-coil motif-containing protein [GOPC], Bromodomain-containing protein 4 [BRD4], and androgen receptor [AR]), with protein ESR1 having the most peptide interaction of 112 peptides, followed by GOPC (48), BRD4 (8), and AR (1).

Prediction using model 11 produced 765 peptide-protein pairs with 140 distinct peptides and 66 proteins. The top 5 protein with the most interaction with a peptide is ESR1 with 90 peptides, followed by EGFR (76), GOPC (69), BCL6 (47), and YWHAE/14-3-3 protein epsilon (42).

Venom peptide prediction with these three models also yielded the same 121 pairs of peptide-protein that consisted of 86 unique peptides and 3 proteins (ESR1, GOPC, and BRD4). Supplementary File 2 shows all of these overlap prediction results.

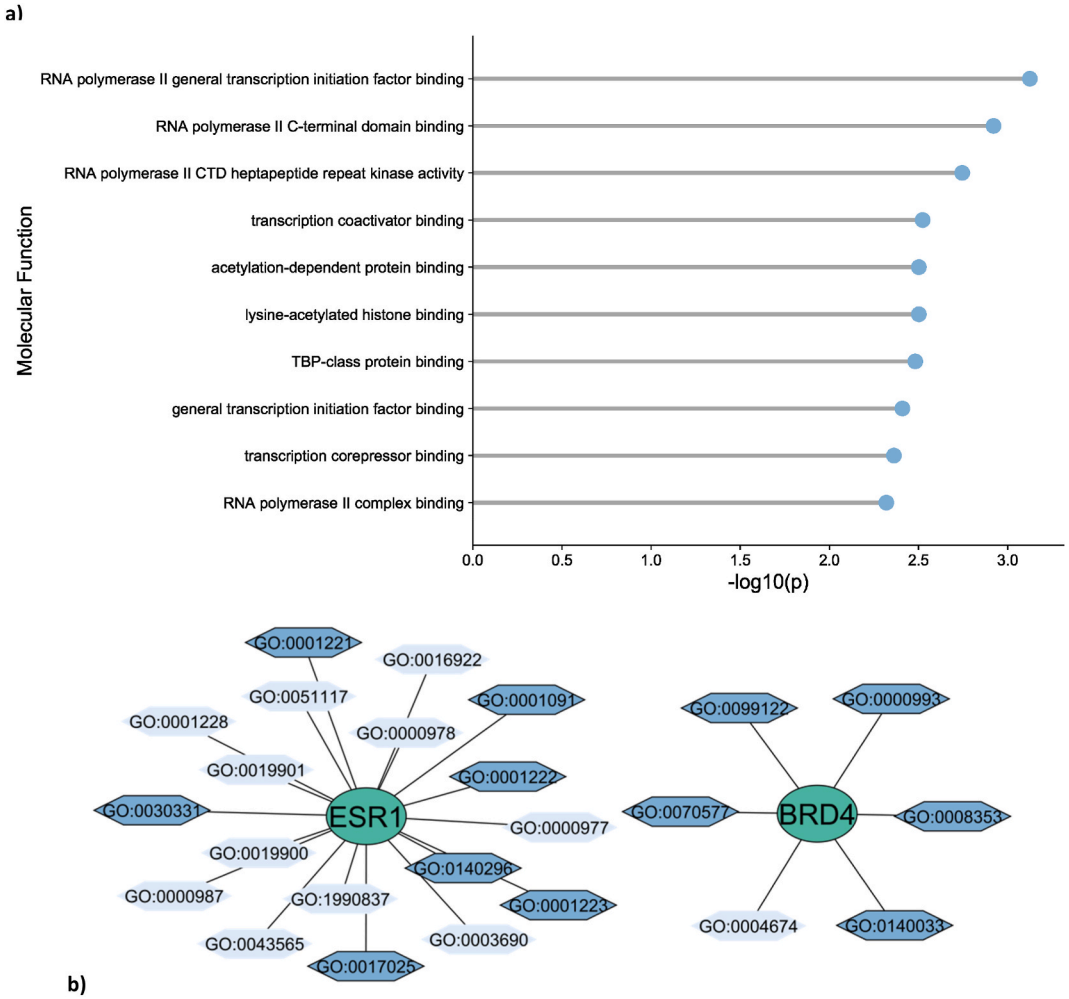


Fig. 8. GO Molecular Function: a. Top 10 terms with the lowest *p*-value; b. network of protein-GO Molecular Function, darker colors indicate the terms with a *p*-value of <0.01 (enrichment terms explanation can be seen in Supplementary File 5). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

3.4. Venom peptide prediction results in dataset version 2

Venom peptide prediction using dataset version 2 produced a slight difference result. Model 5 prediction resulted in 145 pairs of peptide-protein with all peptides predicted to have interaction with only 1 protein (CREBPP).

Model 10 prediction yielded 121 pairs of peptide-protein with 108 unique peptides predicted to have an interaction with only three proteins (ESR1, GOPC, and BRD4), with protein ESR1 having the most peptide interaction of 107 peptides, followed by GOPC (7) and BRD4 (7).

Prediction using model 11 produced 1056 peptide-protein pairs with 140 distinct peptides and 71 proteins. The top 5 protein with the most interaction with a peptide is ESR1 with 102 peptides, followed by AR (86), BRD4 (83), GOPC (73), and BCL6 (58). Peptide-protein pairs have no overlap among these three prediction model results. The overlap pairs were found in models 10 and 11 prediction results, with a total of 95 pairs that consisted of 86 unique peptides and 3 proteins (ESR1, GOPC, and BRD4). Supplementary File 3 shows all of these overlap prediction results.

A similarity was found in pairs of peptide-protein between the prediction of dataset versions 1 and 2, with 67 pairs which consisted of 64 unique peptides and 3 proteins (ESR1, GOPC, and BRD4). Supplementary File 4 shows these overlap pairs.

3.5. Enrichment analysis results

The identification results of protein overlap interactions with peptides included 83 peptides for ESR1, 6 for GOPC, and 6 for BRD4. Gene ontology and KEGG pathways are identified from ESR1, GOPC, and BRD4 proteins using the Enrichr database, as seen in Figs. 6–9 and Supplementary File 5. Fig. 10 shows the topological network of protein-peptide-enrichment.

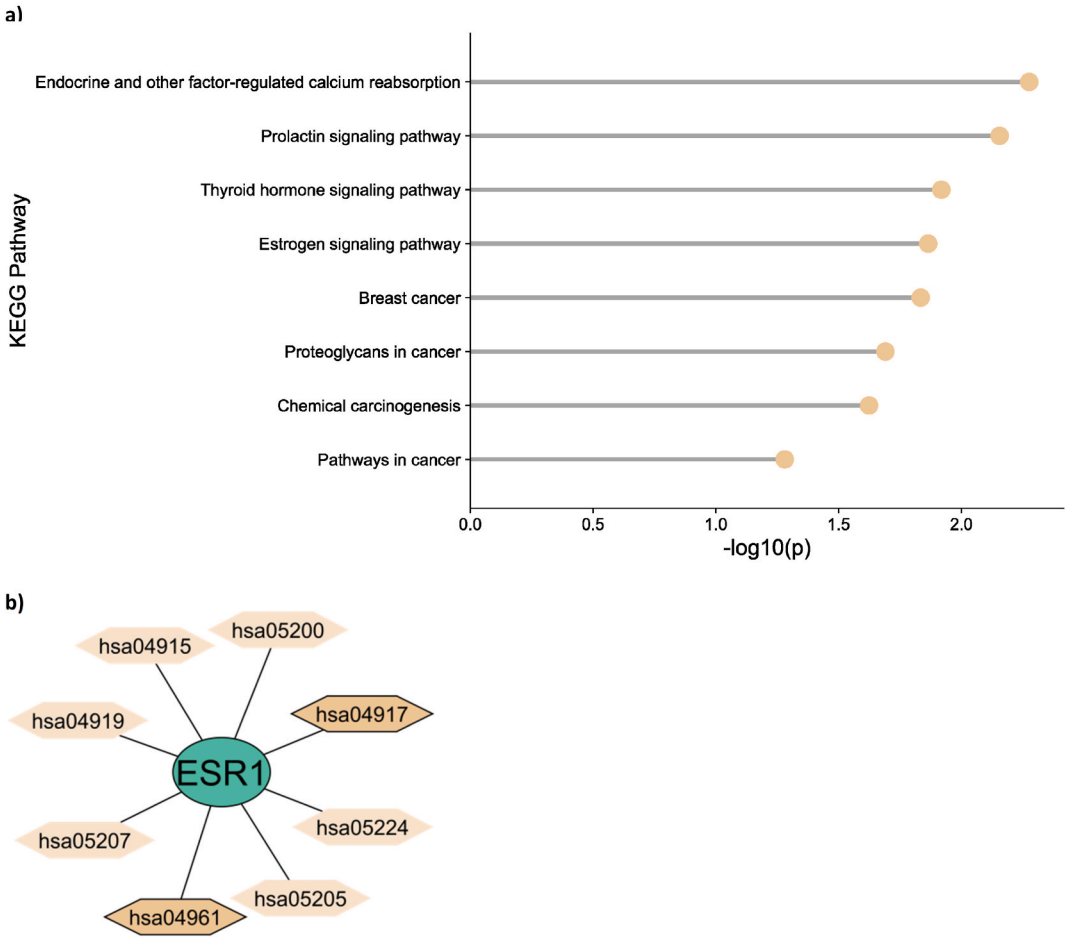


Fig. 9. KEGG Pathway: a. Top 10 terms with the lowest *p*-value; b. network of protein-KEGG Pathway, darker colors indicate the terms with a *p*-value of <0.01 (enrichment terms explanation can be seen in Supplementary File 5). (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

Protein selection was based on the appearance of the three proteins from the overlapping venom peptide prediction results of dataset versions 1 and 2. The pathway enrichment analysis results revealed 121 GO terms consisting of 83 biological processes (Figs. 6b), 14 cellular components (Figs. 7b), and 24 molecular functions (Fig. 8b). The identified KEGG pathways included eight pathways with ESR1 as the associated protein (Fig. 9b). The top 10 terms with the lowest p-value from each gene ontology and KEGG pathway can be seen in Fig. 6a (biological processes), 7a (cellular components), 8a (molecular functions), and 9a (KEGG pathways).

The analysis revealed the five lowest *p*-values of biological processes terms consist of regulation of I-kappaB kinase/NF-kappaB signaling (ESR1 and BRD4), negative regulation of anion transmembrane transport (GOPC), regulation of anion channel activity (GOPC), positive regulation of RNA polymerase II transcription preinitiation complex assembly (ESR1), and apical protein localization (GOPC). The GOPC protein revealed no molecular function results. The five lowest *p*-values of molecular function terms were RNA polymerase II general transcription initiation factor binding (ESR1), RNA polymerase II C-terminal domain binding (BRD4), RNA polymerase II CTD heptapeptide repeat kinase activity (BRD4), transcription coactivator binding (ESR1), and acetylation-dependent protein binding (BRD4). The cellular component result with the five lowest *p*-values was trans-Golgi network transport wascle (GOPC), condensed nuclear chromosome (BRD4), Golgi-associated vesicle membrane (GOPC), Golgi-associated vesicle (GOPC), and condensed chromosome (BRD4).

ESR1, GOPC, and BRD4 are proteins involved in cancer. ESR1 is a gene that codes for estrogen receptor α (ER-a) [30]. ER-α is a ligand-activated transcription factor that regulates the transcription of estrogen-sensitive genes [31]. ESR1 gene mutations can cause estrogen dysfunction and increase the risk of breast cancer [32]. ESR1 dysregulation also increases breast cancer's metastatic potential and therapeutic resistance [33], which causes ESR1 as an important prognostic factor [34]. The role of ESR1 in cancer is seen from the KEGG pathway results, including pathways in cancer, breast cancer, proteoglycans in cancer, chemical carcinogenesis, and thyroid hormone signaling pathways, estrogen signaling pathways, prolactin signaling, and endocrine and other factor-regulated calcium reabsorption.

GOPC is localized in the trans-Golgi network, interacting through its single SDF-95/discs large/ZO-1 (PDZ) domain with various

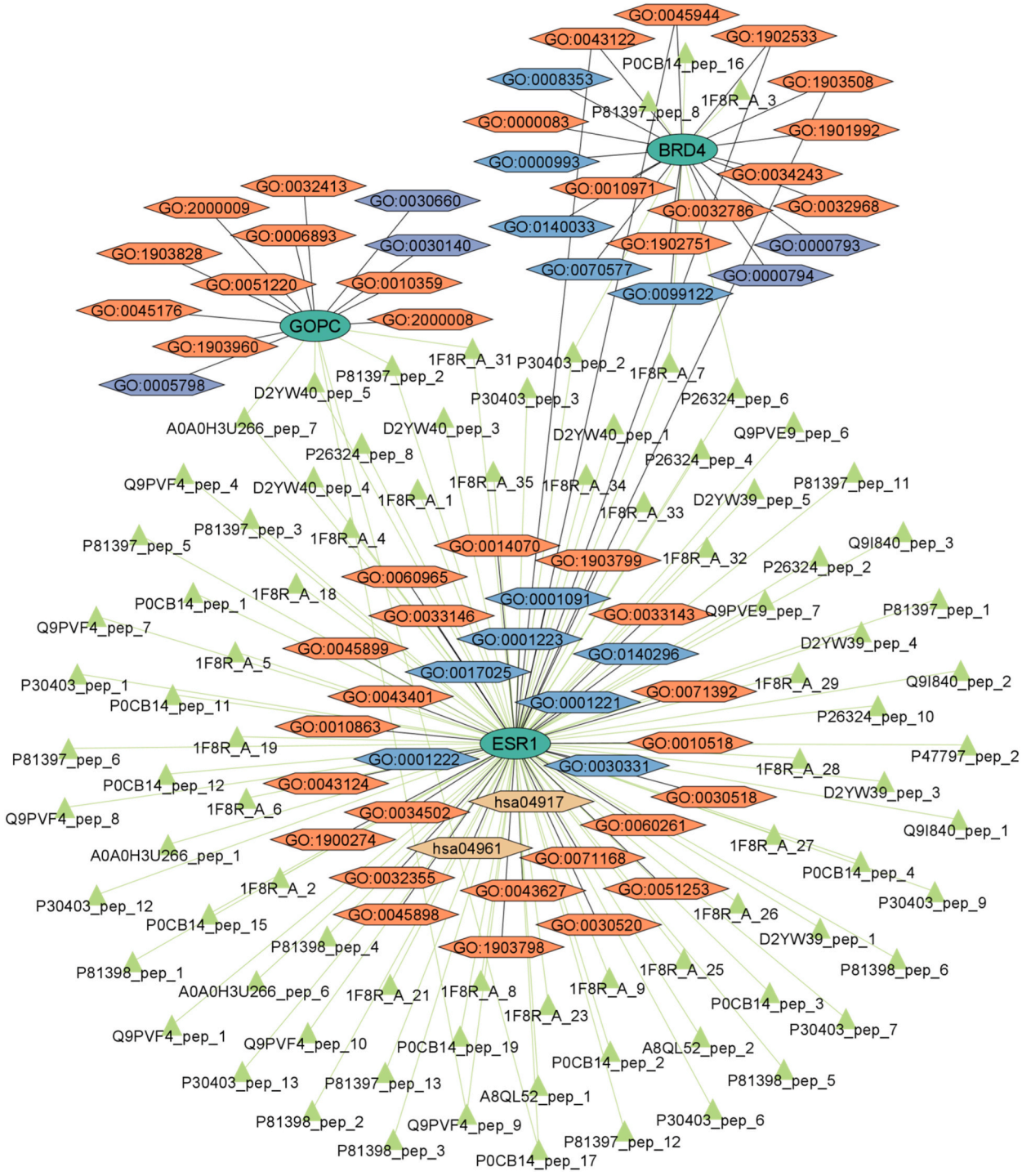


Fig. 10. Topology of protein-peptide-enrichment network: round, triangular, and hexagonal shaped nodes indicate protein, peptide, and enrichment; the enrichment terms have a p-value of <0.01 (terms explanation can be seen in Supplementary File 1–5).

cell surface receptors [35,36]. GOPC controls the intercellular trafficking of numerous integral membrane proteins. Lower expression of GOPC correlates with a worse prognosis in patients with colorectal cancer [37].

BRD4 is a member of the Bromodomain and Extra-terminal (BET) protein family, which is a cancer protein because it plays a role in super-enhancers organization and regulation of cancer gene expression [38]. Additionally, BET also plays a vital role as a therapeutic target in diseases, such as inflammation, neurological disorders, obesity [39], and cardiovascular [40]. Meanwhile, BRD4 is an

epigenetic reader that binds to acetylated histones, regulates gene transcription and proliferation, and repairs DNA damage. BRD4 can trigger tumor growth and inflammation [41–43]. BRD4 hyperphosphorylation contributes to bromodomain and extra-terminal inhibitor (BETi) resistance, where BRD4 phosphorylation supports chromatin binding and higher oncogene expression [42].

4. Conclusions

This study conducted the prediction of peptide-protein interactions in cancer cases using SAE-DNN, RF, and XGBoost. The results revealed that all algorithms produced fairly similar results, with the XGBoost model (scenario 11) considered the best model among other algorithms and scenarios due to its good results in accuracy (0.859), precision (0.663), ROC-AUC (0.697), and the number of true positive and false positive in its prediction results. Enrichment analysis of the three proteins (ESR1, GOPC, and BRD4) overlapping from models 5, 10, and 11 showed that the proteins are related to cancer disease. The KEGG Pathway analysis revealed that ESR1 protein has associated with several pathways related to cancer, including pathways in cancer (hsa05200), chemical carcinogenesis (hsa05207), breast cancer (hsa05224), and proteoglycans in cancer (hsa05205).

The methods proposed in this study to predict peptide-protein interaction in cancer cases have several advantages. Firstly, the methods only require sequence information on peptide and protein as an input, which can speed up the training process instead of using structure data of the peptide and protein. Moreover, comparing multiple methods of machine learning and deep learning models can help narrow down the best models to predict the interaction, as well as enhance the prediction results through the overlap interaction that occurred in the models.

However, this study still has several limitations. Firstly, the amount of data used is considered small due to the small number of samples (peptide and protein) that have been acquired for cancer cases. This study did not produce an overall good performance due to its low results in recall and Fscore. Additionally, the prediction in this study still cannot predict the binding sites and peptide-binding residue involved in the interactions. Therefore, more data should be collected regarding the interaction of peptides in cancer proteins and more advanced algorithms should be used. Predicting the binding residue in each interaction should be attempted to improve the prediction results.

Author contribution statement

Wisnu Ananta Ananta Kusuma: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper.

Aulia Fadli: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Rizka Fatriani: Performed the experiments; Analyzed and interpreted the data; Wrote the paper.

Tri Rini Nuringtyas; Donan Satria Yudha; Respati Tri Swasono: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Fajar Sofyantoro: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Kenny Lischer; Wahyu Aristyaning Putri; Yekti Asih Purwestri: Contributed reagents, materials, analysis tools or data.

Data availability statement

Data will be made available on request.

Funding

This research and APC was funded by IPB University, Indonesia, under the Indonesian Collaborative Research (Riset Kolaborasi Indonesia) 2022 grant number 3334/IT3.L1/PT.01.03/P/B/2022.

Declaration of competing interest

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

Acknowledgments

We thank the IPB University for providing research funding under the Indonesian Collaborative Research (Riset Kolaborasi Indonesia) 2022.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e21149.

References

[1] B.M. Cooper, J. Iegre, D.H. O'Donovan, M. Ölwegård Halvarsson, D.R. Spring, Peptides as a platform for targeted therapeutics for cancer: peptide-drug conjugates (PDCs), Chem. Soc. Rev. 50 (2021) 1480–1494, https://doi.org/10.1039/d0cs00556h.

[2] M. Alas, A. Saghaeidehkordi, K. Kaur, Peptide-drug conjugates with different linkers for cancer therapy, J. Med. Chem. 64 (2021) 216–232, https://doi.org/10.1021/acs.jmedchem.0c01530.

[3] A.C.L. Lee, J.L. Harris, K.K. Khanna, J.H. Hong, A comprehensive review on current advances in peptide drug development and design, Int. J. Mol. Sci. 20 (2019) 2383, https://doi.org/10.3390/ijms20102383.

[4] Y. Lei, S. Li, Z. Liu, F. Wan, T. Tian, S. Li, D. Zhao, J. Zeng, A deep-learning framework for multi-level peptide–protein interaction prediction, Nat. Commun. 12 (2021), https://doi.org/10.1038/s41467-021-25772-4.

[5] A. Cichonska, B. Ravikumar, E. Parri, S. Timonen, T. Pahikkala, A. Airola, K. Wennerberg, J. Rousu, T. Aittokallio, Computational-experimental approach to drug-target interaction mapping: a case study on kinase inhibitors, PLoS Comput. Biol. 13 (2017), https://doi.org/10.1371/journal.pcbi.1005678.

[6] Y. Liu, M. Wu, C. Miao, P. Zhao, X.L. Li, Neighborhood regularized logistic matrix factorization for drug-target interaction prediction, PLoS Comput. Biol. 12 (2016), 1004760, https://doi.org/10.1371/journal.pcbi.1004760.

[7] N.S. Ramadhanti, W.A. Kusuma, I. Batubara, R. Heryanto, Random forest to predict eucalyptus as a potential herb in preventing coviD19, 2021, IEEE Conf. Comput. Intell. Bioinforma. Comput. Biol. CIBCB 2021 (2021) 1–5, https://doi.org/10.1109/CIBCB49929.2021.9562940.

[8] S.Z. Sajadi, M.A. Zare Chahooki, S. Gharaghani, K. Abbasi, AutoDTI++: deep unsupervised learning for DTI prediction by autoencoders, BMC Bioinf. 22 (2021) 1–19, https://doi.org/10.1186/S12859-021-04127-2/FIGURES/5.

[9] F. Sulistiawan, W.A. Kusuma, N.S. Ramadhanti, A. Tedjo, Drug-target interaction prediction in coronavirus disease 2019 case using deep semi-supervised learning model, in: 2020 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2020, Institute of Electrical and Electronics Engineers Inc., 2020, pp. 83–88, https://doi.org/10.1109/ICACSIS51025.2020.9263241.

[10] A. Fadli, W.A. Kusuma, Annisa, I. Batubara, R. Heryanto, Screening of potential Indonesia herbal compounds based on multi-label classification for 2019 coronavirus disease, Big Data Cogn, Comput. Times 5 (2021) 75, https://doi.org/10.3390/bdcc5040075.

[11] D.J. Dittman, T.M. Khoshgoftaar, A. Napolitano, The effect of data sampling when using random forest on imbalanced bioinformatics data, in: Proc. - 2015 IEEE 16th Int. Conf. Inf. Reuse Integr. IRI 2015, Institute of Electrical and Electronics Engineers Inc., 2015, pp. 457–463, https://doi.org/10.1109/IRI.2015.76.

[12] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444, https://doi.org/10.1038/nature14539.

[13] W. Zheng, L. Yang, R.J. Genco, J. Wactawski-Wende, M. Buck, Y. Sun, SENSE: siamese neural network for sequence embedding and alignment-free comparison, Bioinformatics 35 (2019) 1820–1828, https://doi.org/10.1093/bioinformatics/bty887.

[14] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The protein Data Bank, Nucleic Acids Res. 28 (2000) 235–242, https://doi.org/10.1093/nar/28.1.235.

[15] D.S. Wishart, C. Knox, A.C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, J. Woolsey, DrugBank: a comprehensive resource for in silico drug discovery and exploration, Nucleic Acids Res. 34 (2006) D668–D672, https://doi.org/10.1093/nar/gkj067.

[16] S. Salentin, S. Schreiber, V.J. Haupt, M.F. Adasme, M. Schroeder, PLIP: fully automated protein-ligand interaction profiler, Nucleic Acids Res. 43 (2015) W443–W447, https://doi.org/10.1093/nar/gkv315.

[17] A.L. Lehninger, D.L. Nelson, M.M. Cox, Principles of Biochemistry: Advance Chapters from the 2000 Edition, Worth Publishers, 1999.

[18] B. Mészáros, G. Erdös, Z. Dosztányi, IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding, Nucleic Acids Res. 46 (2018) W329–W337, https://doi.org/10.1093/nar/gky384.

[19] F. Madeira, Y.M. Park, J. Lee, N. Buso, T. Gur, N. Madhusoodanan, P. Basutkar, A.R.N. Tivey, S.C. Potter, R.D. Finn, R. Lopez, The EMBL-EBI search and sequence analysis tools APIs in 2019, Nucleic Acids Res. 47 (2019) W636–W641, https://doi.org/10.1093/nar/gkz268.

[20] T. Hamp, B. Rost, Evolutionary profiles improve protein-protein interaction prediction from sequence, Bioinformatics 31 (2015) 1945–1950, https://doi.org/10.1093/bioinformatics/btv077.

[21] X. Liu, L. Zhao, Q. Dong, Protein remote homology detection based on auto-cross covariance transformation, Comput. Biol. Med. 41 (2011) 640–647, https://doi.org/10.1016/j.compbiomed.2011.05.015.

[22] S. Redkar, S. Mondal, A. Joseph, K.S. Hareesha, A machine learning approach for drug-target interaction prediction using wrapper feature selection and class balancing, Mol. Inform. 39 (2020), 1900062, https://doi.org/10.1002/minf.201900062.

[23] M. Bahi, M. Batouche, Drug-target interaction prediction in drug repositioning based on deep semi-supervised learning, in: IFIP Adv. Inf. Commun. Technol., Springer New York LLC, 2018, pp. 302–313, https://doi.org/10.1007/978-3-319-89743-1_27.

[24] Y.L. Pavlov, Random Forests, Springer, 2019, https://doi.org/10.4324/9781003109396-5.

[25] Y. Qi, Random forest for bioinformatics, in: Ensemble Mach. Learn., Springer US, 2012, pp. 307–323, https://doi.org/10.1007/978-1-4419-9326-7_11.

[26] B. Yu, W. Qiu, C. Chen, A. Ma, J. Jiang, H. Zhou, Q. Ma, SubMito-Xgboost, Predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting, Bioinformatics 36 (2020) 1074–1081, https://doi.org/10.1093/bioinformatics/btz734.

[27] K.H. Brodersen, C.S. Ong, K.E. Stephany, J.M. Buhmann, The binormal assumption on precision-recall curves, in: Proc. - Int. Conf. Pattern Recognit., 2010, pp. 4263–4266, https://doi.org/10.1109/ICPR.2010.1036.

[28] E.Y. Chen, C.M. Tan, Y. Kou, Q. Duan, Z. Wang, G.V. Meirelles, N.R. Clark, A. Ma'ayan, Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool, BMC Bioinf. 14 (2013), https://doi.org/10.1186/1471-2105-14-128.

[29] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504, https://doi.org/10.1101/gr.1239303.

[30] D.K. Lung, R.M. Reese, E.T. Alarid, Intrinsic and extrinsic factors governing the transcriptional regulation of ESR1, Horm. Cancer. 11 (2020) 129–147, https://doi.org/10.1007/s12672-020-00388-0.

[31] M.T. Pagano, E. Ortona, M.L. Dupuis, A role for estrogen receptor alpha36 in cancer progression, Front. Endocrinol. 11 (2020) 1–7, https://doi.org/10.3389/fendo.2020.00506.

[32] M. Saneipour, A. Sheikhi, A. Moridnia, An interdependence between estrogen receptor 1 gene polymorphisms and susceptibility to breast cancer, Trends Med. Sci. 1 (2021) 1–7, https://doi.org/10.5812/tms.117221.

[33] J.T. Lei, X. Gou, S. Seker, M.J. Ellis, ESR1 alterations and metastasis in estrogen receptor positive breast cancer, J. Cancer Metastasis Treat. 2019 (2019), https://doi.org/10.20517/2394-4722.2019.12.

[34] H. Liao, W. Huang, W. Pei, H. Li, Detection of ESR1 mutations based on liquid biopsy in estrogen receptor-positive metastatic breast cancer: clinical impacts and prospects, Front. Oncol. 10 (2020) 1–9, https://doi.org/10.3389/fonc.2020.587671.

[35] J. Koliwer, M. Park, C. Bauch, M. Von Zastrow, H.J. Kreienkamp, The golgi-associated PDZ domain protein PIST/GOPC stabilizes the β1-Adrenergic receptor in intracellular compartments after internalization, J. Biol. Chem. 290 (2015) 6120–6129, https://doi.org/10.1074/jbc.M114.605725.

[36] M. Klüssendorf, I. Song, L. Schau, F. Morellini, A. Dityatev, J. Koliwer, H.J. Kreienkamp, The golgi-associated PDZ domain protein gopc/PIST is required for synaptic targeting of mGluR5, Mol. Neurobiol. 58 (2021) 5618–5634, https://doi.org/10.1007/s12035-021-02504-9.

[37] N. Ohara, N. Haraguchi, J. Koseki, Y. Nishizawa, K. Kawai, H. Takahashi, J. Nishimura, T. Hata, T. Mizushima, H. Yamamoto, H. Ishii, Y. Doki, M. Mori, Low expression of the GOPC is a poor prognostic marker in colorectal cancer, Oncol. Lett. 14 (2017) 4483–4490, https://doi.org/10.3892/ol.2017.6817.

[38] B. Donati, E. Lorenzini, A. Ciarrocchi, BRD4 and Cancer: going beyond transcriptional regulation, Mol. Cancer 17 (2018) 1–13, https://doi.org/10.1186/s12943-018-0915-9.

[39] B. Padmanabhan, S. Mathur, R. Manjula, S. Tripathi, Bromodomain and extra-terminal (BET) family proteins: new therapeutic targets in major diseases, J. Bio. Sci. 41 (2016) 295–311, https://doi.org/10.1007/s12038-016-9600-6.

[40] S.Y. Kim, X. Zhang, G.G. Schiattarella, F. Altamirano, T.A.R. Ramos, K.M. French, N. Jiang, P.A. Szweda, B.M. Evers, H.I. May, X. Luo, H. Li, L.I. Szweda, V. Maracaja-Coutinho, S. Lavandero, T.G. Gillette, J.A. Hill, Epigenetic reader BRD4 (Bromodomain-Containing protein 4) governs nucleus-encoded mitochondrial transcriptome to regulate cardiac function, Circulation 142 (2020) 2356–2370, https://doi.org/10.1161/CIRCULATIONAHA.120.047239.
[41] Z. Zhou, X. Li, Z. Liu, L. Huang, Y. Yao, L. Li, J. Chen, R. Zhang, J. Zhou, L. Wang, Q.Q. Zhang, A bromodomain-containing protein 4 (BRD4) inhibitor suppresses angiogenesis by regulating AP-1 expression, Front. Pharmacol. 11 (2020) 1–11, https://doi.org/10.3389/fphar.2020.01043.
[42] R. Wang, J.F. Yang, F. Ho, E.S. Robertson, J. You, Bromodomain-containing protein BRD4 is hyperphosphorylated in mitosis, Cancers 12 (2020) 1–20, https://doi.org/10.3390/cancers12061637.
[43] A.L. Drumond-Bock, M. Bieniasz, The role of distinct BRD4 isoforms and their contribution to high-grade serous ovarian carcinoma pathogenesis, Mol. Cancer 20 (2021) 1–15, https://doi.org/10.1186/s12943-021-01424-5.