Research article

Identification of Repurposable Drugs for Colorectal Cancer Using Drug-Network-Based Classification Models

Keeratika Thongchaiprasit[1], Natthakarn Ariyasajjakorn[1], Nichapa Chatjindarat[1], Sittichoke Som-am[1] and Thitipong Kawichai[2]*

[1]*Mahidol Wittayanusorn School, Nakhon Pathom, Thailand*
[2]*Department of Mathematics and Computer Science, Academic Division, Chulachomklao Royal Military Academy, Nakhon Nayok, Thailand*

Curr. Appl. Sci. Technol. 2024, Vol. 24 (No. 3), e0258109; https://doi.org/10.55003/cast.2024.258109

Received: 25 April 2023, Revised: 28 July 2023, Accepted: 15 November 2023, Published: 15 February 2024

Keywords

classification;
colorectal cancer;
drug repurposing;
multi-layer perceptron;
networks

Abstract

Colorectal cancer (CRC) is the second most lethal cancer with more than one million new cases diagnosed worldwide every year. To defuse the increasing CRC threat, more effective and less harmful treatments for CRC patients are urgently needed. Computational drug repurposing, which is an in silico based approach to uncover new indications of approved drugs, is a promising strategy to accelerate the time to market of drugs. However, there are not many computational drug repurposing methods for CRC. In this work, we proposed drug-network-based classification models to identify repurposable drugs for CRC. Initially, four drug networks, the chemical structure network (CSN), the target protein network (TPN), the drug pathway network (PWN), and the drug-drug interaction network (DIN), were formulated. Based on the drug features properly extracted from the networks, we created four multi-layer perceptron (MLP) models. By comparing the performance of the models, the DIN model outperformed the others with the highest accuracy and an F_1 score of 96.9%. After predicting the repurposability of over 1,200 non-CRC approved drugs using the DIN model, 306 drugs discovered as potentially repurposable drugs for CRC. In summary, the drug-network-based classification models can efficiently identify repurposable drug candidates for CRC, which would be applicable for efficient therapeutic treatment of CRC.

1. Introduction

Colorectal cancer (CRC) is one of the most common cancer types found in both males and females. More than one million new CRC cases are diagnosed worldwide every year [1]. Furthermore, CRC is the second most fatal cancer with over 900,000 global deaths reported in 2020 [2]. The prevalence and

*Corresponding author: Tel.: (+66) 37393010 ext. 62292
E-mail: thitipong.kaw@crma.ac.th

mortality rates of CRC have tended to rise every year due to present-day lifestyles and behaviors of food consumption, such as eating only low-fiber foods and lack of physical exercise [3]. It was estimated that the numbers of new CRC cases and deaths will increase to 3.2 million cases and 1.6 deaths per year by 2040 [4]. Thus, new approaches or drugs which are more effective and less harmful for the treatment of CRC are urgently required to defuse the CRC threat in the future.

Drug repurposing, which is about uncovering new indications of known drugs that have been approved by the Food and Drug Administration (FDA), is a promising approach to shorten drug discovery timelines and accelerate the time to market of drugs. Studies on drug repurposing for CRC can be broadly divided into two groups, including an experimental approach and a computational approach. Experimental approaches featuring *in vitro* or *in vivo* models have been used to uncover many potentially repurposable drug candidates for CRC treatment. For example, lovastatin, a drug originally indicated for lipid reduction, was revealed to suppress CRC metastasis formation via a high-throughput screening [5]. Moreover, an anti-retroviral drug named efavirenz was found to activate the phosphorylation of tumor protein p53, leading to apoptosis of CRC cell lines [6]. Nevertheless, drug repurposing with experimental wet labs still requires a lot of materials and resources, particularly for the large-scale screening of compounds.

To support large-scale drug repurposing, various computational methods have been developed, and they are generally based on data mining, machine learning, and network analysis. Since the number of drugs approved for diseases such as CRC is still limited, there are currently not many computational drug repurposing methods developed for CRC. There are few methods created for identifying potential drug targets of CRC. Tao *et al*. [7] proposed an ontology-based method to predict potential CRC drug targets that provided valuable information for the development of anti-CRC drugs. Irham *et al*. [8] utilized text mining and network analysis to study genetic variants of CRC and identify promising new targets for further discovering CRC repurposable drugs. In case of drug repurposing, most existing methods for generic identification of repurposable drugs using integrated data involved with multiple diseases were proposed. For example, a three-layer heterogeneous network model (TL-HGBI) [9] and a method of bi-random walks (MBiRW) [10] integrated data from numerous drugs and diseases as heterogeneous networks for identifying drug-disease associations. Numerous association data among drugs, diseases, proteins, and gene ontology (GO) were utilized in the methods of topological similarity and singular value decomposition (TS-SVD) [11] and meta-path based gene ontology profiles in order to predict drug-disease associations (MGP-DDA) [12]. Over the last few years, several methods have been developed based on various deep learning models such as a model for potential drug-disease interactions prediction (DDIPred) [13] and a method for identifying drug-disease associations based on a geometric deep learning framework (DDAGDL) [14]. Although there are many existing methods for generic drug repurposing, deploying those methods in identifying repurposable drugs for a particular disease is often difficult due to numerous training data required and challenges in implementation.

In this work, we proposed drug-network-based classification models to discover potentially repurposable drugs for CRC. To develop the models, a drug chemical structure network (CSN), a drug target protein network (TPN), a drug pathway network (PWN), and a drug-drug interaction network (DIN) were initially constructed. To solve the data limitation of CRC approved drugs, we applied the synthetic minority oversampling (SMOTE) method. From each drug network, drug features were prepared and utilized to create a classification model of multi-layer perceptron (MLP). By comparing the performance of all drug-network-based models, the best classification model was selected and used to predict the repurposing ability of approved drugs for CRC.

2. Materials and Methods

An overview of this work is illustrated in Figure 1. Initially, the drug data including chemical structures, target proteins, drug pathways, and drug-drug interactions were collected from numerous databases. The relationships between drugs and their properties corresponding to each data set were formulated as a drug network. Consequently, there were four drug networks: the drug chemical structure network (CSN), the drug target protein network (TPN), the drug pathway network (PWN), and the drug-drug interaction network (DIN). From each drug network, different drug features were extracted and prepared by using Multiple Correspondence Analysis (MCA). Then, we used the drug features and the synthetic minority oversampling method (SMOTE) to create a multi-layer perceptron (MLP) model of each drug network. All drug-network-based models were evaluated their performance, and then the best model was chosen to finally predict the repurposing ability of some unknown drugs.

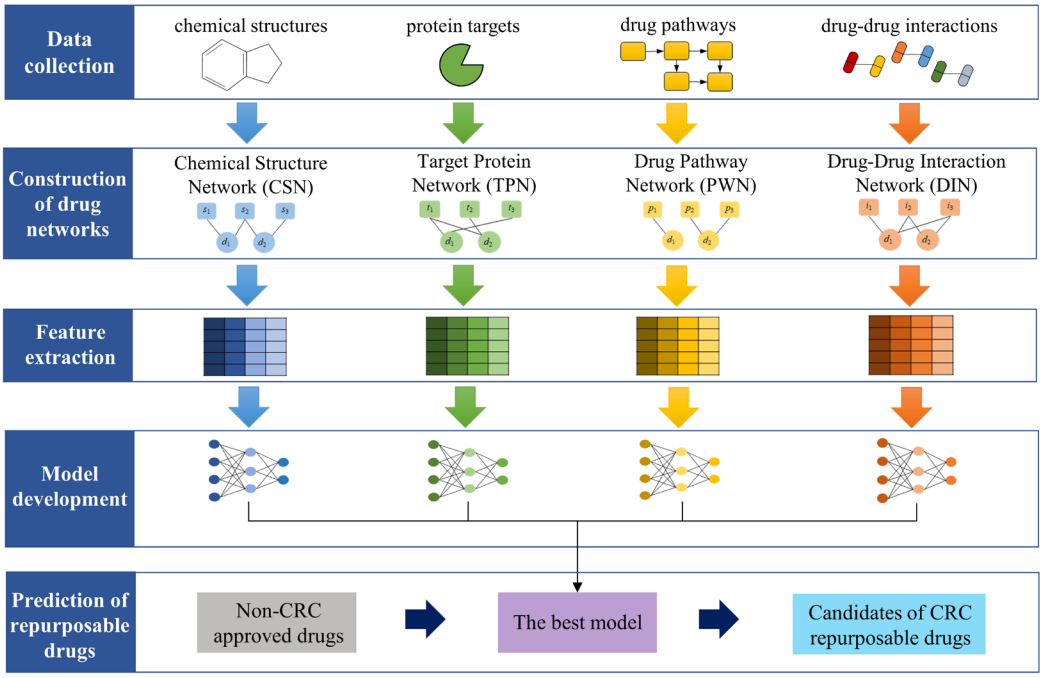

Figure 1. An overview of this study

2.1 Data collection and preparation

Initially, a list of all FDA approved drugs was downloaded from DrugBank [15]. In this list, each drug was noted if it was approved for the treatment of CRC or not. Next, several properties of the approved drugs, including chemical structures, drug protein targets, drug pathways, and drug-drug interactions, were collected from numerous databases. Drug chemical structures in the form of the simplified molecular-input line-entry system (SMILES) were downloaded from DrugBank [15], PubChem [16], and ChEMBL [17]. Drug target proteins were collected from DrugBank [15], PubChem [16], ChEMBL [17], and KEGG [18]. Pathways relevant to each approved drug were

mainly collected from PathBank [19] and KEGG [18]. Finally, drug-drug interactions were downloaded from DDInter [20] and KEGG [18].

From DrugBank, we obtained 29 CRC approved drugs and 4,226 non-CRC approved drugs. In these data sets, some drugs did not have chemical structure, target protein, drug pathway, or drug-drug interaction information. Therefore, the non-CRC drugs that have complete information were used for training and testing models, and the remaining non-CRC drugs were used for discovering potentially repurposable drugs for CRC. Due to a very small number of CRC approved drugs, all of them were kept for developing models. In total, there were 321 non-CRC approved drugs for training and testing models.

2.2 Construction of drug networks

A network (or a graph) is a mathematical tool used to represent relationships between single-type or multi-type objects. Basic definitions about networks are provided in definitions 1 to 3.

Definition 1 (Network). A network or a graph G consists of a collection of vertices or nodes V and a collection of edges E where each edge $e \in E$ joins two vertices in V. We can write $G = (V, E)$ to represent the network.

Definition 2 (Bipartite network). A network $G = (V, E)$ is a bipartite network if V can be partitioned into two disjoint sets V_1 and V_2. Additionally, each edge $e \in E$ links $v_1 \in V_1$ to $v_2 \in V_2$, or $E \subseteq \{\langle v_1, v_2 \rangle \mid v_1 \in V_1, v_2 \in V_2\}$. A bipartite network is also called a two-mode network [21].

Definition 3 (Incidence matrix). Let $G = (V, E)$ be a bipartite network of two disjoint sets of vertices V_1 and V_2. $n(V_1)$ and $n(V_2)$ are the number of elements in V_1 and V_2, respectively. An incidence matrix A is an $n(V_1) \times n(V_2)$ matrix where an element $A_{ij} = 1$ if an edge $e \in E$ links vertex i to j and $A_{ij} = 0$ otherwise [22].

Based on the collected data sets, we created four drug networks including the drug chemical structure network (CSN), the target protein network (TPN), the drug pathway network (PWN), and the drug-drug interaction network (DIN), to represent different information of the approved drugs. Examples of the drug networks and their incidence matrices are demonstrated in Figure 2. All drug networks created in this work were bipartite networks. In Figure 2, nodes *d*, *s*, *t*, *p*, and *i* represent nodes of DrugBank approved drugs, chemical substructures of drugs, drug target proteins, PathBank pathways, and DDInter drugs interacting with DrugBank drugs, respectively. In an incidence matrix, 1 represents an existing edge between a pair of nodes, and 0 indicates an absence of edges.

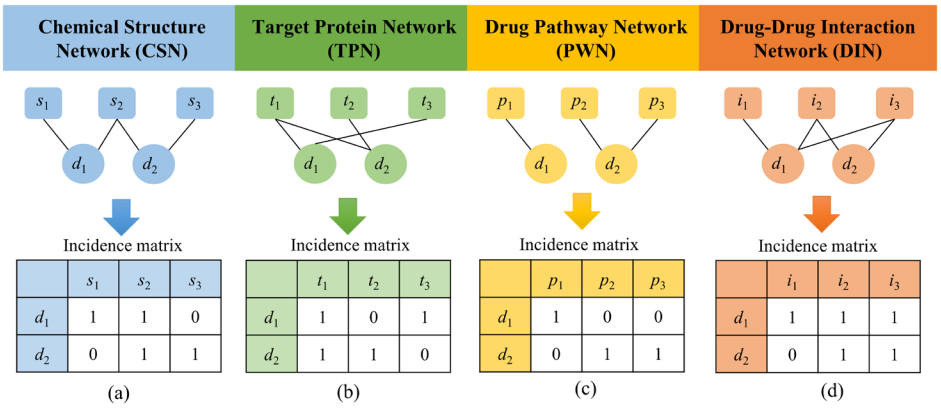

Figure 2. Examples of all drug networks and their incidence matrices

To construct CSN, we firstly used the RDKit library in Python to generate 1024-bit vectors of the Morgan fingerprints of the approved drugs. Each fingerprint bit indicates the existence of a chemical substructure of the drug. Based on those chemical fingerprints, we created a bipartite network of drug chemical structures, or CSN, which consisted of nodes of the approved drugs and components of the Morgan fingerprints. In the CSN, edges connected nodes of drugs and their chemical fragments, as shown in Figure 2(a). The dimension of the CSN incidence matrix was n_{AD} by 1024, where n_{AD} was the number of the approved drugs. From the data of drug target proteins, we created a bipartite network of drug target proteins, termed TPN, consisting of the nodes of drugs and target proteins. In the TPN, edges link nodes of drugs to their interacting target proteins, as illustrated in Figure 2(b). The dimension of the TPN incidence matrix was n_{AD} by n_{TP}, where n_{TP} was the number of all target proteins interacting with the approved drugs. Similar to TPN, a bipartite network of drug pathways or PWN was composed of nodes of drugs and PathBank pathways. In the PWN, drug nodes and all their involved pathways were linked by edges (Figure 2(c)). The PWN incidence matrix had the dimension of $n_{AD} \times n_{PW}$, where n_{PW} was the number of all PathBank pathways. From the data of drug-drug interactions, we constructed a bipartite network of drug-drug interactions or DIN. This network consisted of nodes of the DrugBank approved drugs and nodes of DDInter drugs. In the DIN, an edge linked a DrugBank drug to an interacting drug found in the DDInter, as shown in Figure 2(d). The incidence matrix of DIN had the dimension of $n_{AD} \times n_{DI}$, where n_{DI} was the total number of DDInter drugs interacting with the approved drugs.

2.3 Feature extraction from drug networks

In this step, drug features were properly extracted from each drug network for further construction of the classification models. The incidence matrices of all drug networks were used as inputs for the feature extraction. We applied multiple correspondence analysis (MCA) to extract drug features. This method is a technique that is widely used to analyze data with a large number of nominal variables and to reduce the dimension of categorical data [23]. To implement MCA, we used an MCA package version 1.0.3 for Python. To control the number of obtained features by MCA, we set the percentage of data variance preserved by all selected dimensions at the default value or 90%.

2.4 Model development

To predict repurposable drugs for CRC, we built four binary classification models based on the drug features extracted from four drug networks. There were the chemical structure network (CSN) model, the target protein network (TPN) model, the drug pathway network (PWN) model, and the drug-drug interaction network (DIN) model. Among all obtained models, we compared their performance to find the best model for the prediction of repurposable drugs from a collection of non-CRC approved drugs.

In this work, a multi-layer perceptron (MLP) algorithm was utilized to create a binary classification model of each drug network. An MLP model is a type of artificial neural network algorithm that can learn both linear and nonlinear relationships in data. MLP is one of the most popular machine learning algorithms widely used in various applications such as image recognition, medical diagnosis, and recommender systems. The architecture of MLP models that we used in this work is shown in Figure 3.

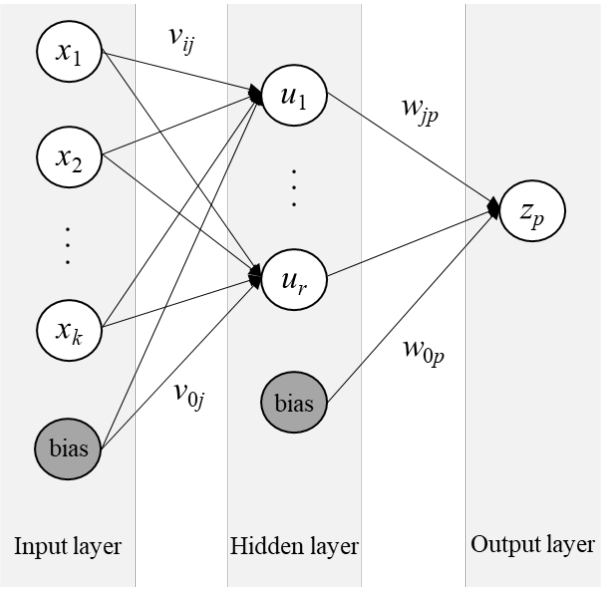

Figure 3. An architecture of MLP models used in this study

From Figure 3, we used a fully connected network of three layers including an input layer, a hidden layer, and an output layer. In the input layer, x_i (i = 1, 2, …, k) was the value of the i^{th} drug feature, and k was the size of the input layer corresponding to the number of all extracted features obtained by MCA. Due to the small amount of training data, we used an MLP model with a single hidden layer. In the hidden layer, a value in a hidden node or u_j (j = 1, 2, …, r) was calculated as shown in equation (1), where r was the number of hidden layer sizes, v_{ij} was the weight for the i^{th} input node and j^{th} hidden node, and v_{0j} was the weight for a bias node. To select a suitable value of r for each drug-network-based model, we varied value r as 10, 20, 30, 40, 50, and investigated the model performance. In the output layer, z_p was the value of the output node calculated by (2), where w_{jp} was the weight for the j^{th} hidden node, w_{0p} was the weight for a bias node, and f was the rectified linear units (ReLU) activation function. A value z_p served as a predicted probability suggesting the repurposing potential of a drug against CRC. To implement the MLP models, we used the scikit-learn module in Python.

$$u_j = \sum_{i=1}^{k} v_{ij} x_i + v_{0j} \tag{1}$$

$$z_p = \sum_{j=1}^{r} w_{jp} f(u_j) + w_{0p} \tag{2}$$

In our data sets, there were a large number of non-CRC approved drugs relative to the number of CRC approved drugs with the approximate ratio of 11:1. To solve this data imbalance problem, we performed the synthetic minority over-sampling technique (SMOTE) for creating synthetic samples of the minority class or CRC approved drugs [24]. Then, we used the balanced data sets obtained from SMOTE in the training and testing of each drug-network-based model. To evaluate the model performance, we conducted 10-fold cross validation. Several standard evaluation

metrics for binary classification, including precision, recall, accuracy, the Matthew's correlation coefficient (MCC), and F_1 score (F_1) were computed, as can be seen in equations (3) - (6). *TP*, *TN*, *FP*, and *FN* were the numbers of true positives, true negatives, false positives, and false negatives, respectively, in the testing data set. In this work, positive and negative samples referred to CRC approved drugs and non-CRC approved drugs, respectively. In addition, we also considered the area under a precision-recall curve (AUPR) and the area under a receiver operating characteristic curve (AUC). These two metrics measure the binary classification ability of a model when the threshold score, or the minimum predicted score of a positive sample, is varied.

$$\text{Precision} = \frac{TP}{TP + FP}, \text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

$$F_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{TP}{TP + \frac{1}{2}(FP + FN)} \tag{6}$$

2.5 Prediction of repurposable drugs

After concluding which drug-network-based model was the best one, we rebuilt that model based on the data of all CRC approved drugs and some of non-CRC approved drugs, without splitting the data into training and testing sets. Due to a very small number of existing CRC drugs, we employed SMOTE for oversampling the minority class, or the class of CRC approved drugs, before reconstructing the model. Using SMOTE, a balanced data set of drugs was obtained and used to rebuild the drug-network-based model. After that, the model was applied to classify the set of all remaining non-CRC approved drugs that were not used to train the model. Non-CRC approved drugs that were predicted to be CRC approved drugs by the model were considered as candidates for repurposable drugs for CRC. To validate these drug candidates, we searched for supporting evidence from published literature, and from ClinicalTrials.gov, which is a publicly available database of clinical studies conducted around the world.

3. Results and Discussion

3.1 Summary of drug network information

In this work, four bipartite networks were constructed based on multiple drug data. The networks were the drug chemical structure network (CSN), the target protein network (TPN), the drug pathway network (PWN), and the drug-drug interaction network (DIN). A summary of information related to all drug networks is shown in Table 1.

Table 1. Summary of drug network information

Drug Network	Type of Node (the number of nodes)	Another Type of Node (the number of nodes)	The Number of Edges
Chemical structure network (CSN)	DrugBank approved drugs (341 nodes)	Components of the Morgan fingerprints (1,024 nodes)	14,529
Drug target protein network (TPN)	DrugBank approved drugs (349 nodes)	Target proteins (1,083 nodes)	4,161
Drug pathway network (PWN)	DrugBank approved drugs (345 nodes)	PathBank pathways (678 nodes)	1,192
Drug-drug interaction network (DIN)	DrugBank approved drugs (343 nodes)	DDInter drugs (1,571 nodes)	53,738

According to Table 1, the CSN consisted of 341 approved drugs and 1,024 Morgan fingerprint components, connected by 14,529 edges in total. The chemical structure of a small molecule is typically composed of approximately 42 Morgan fragments on average. In the TPN, there were 349 approved drugs, 1,083 target proteins, and 4,161 edges linking the drugs and their interacting target proteins. The average degrees of the DrugBank nodes and the protein nodes were about 12 proteins per drug and 4 drugs per protein, respectively. According to the average degrees of nodes in the TPN, it could be suggested that there were many drugs targeting multiple proteins, and there were several proteins typically targeted by multiple drugs. Since the molecular mechanisms of a drug on a disease are related to the proteins that the drug interacts with, the information of many-to-many relationships between drugs and target proteins in the TPN could be useful for the identification of repurposable drugs for CRC.

In the PWN, there were 345 approved drugs, 678 PathBank pathways, and 1,192 edges linking the drugs and their involved pathways. A small molecule is involved in approximately three pathways, and about two drugs are involved in the same pathway on average. Since drug-involved pathways can be used to elucidate the pharmacokinetics of a drug, the information of drug-pathway relationships in the PWN could be used to classify repurposable drugs for CRC. In the DIN, there were 343 DrugBank approved drugs, 1,571 DDInter drugs, and 53,738 edges linking the DrugBank drugs and their interacting DDInter drugs. One approved drug interacted with more than 156 DDInter drugs on average. The larger number of drugs interacting with an approved drug potentially provide much more useful information for the classification of repurposable drugs for CRC.

The incidence matrix of each drug network was used to represent an initial data matrix containing features of drugs based on each drug network. Thus, the numbers of all binary features based on CSN, TPN, PWN, and DIN were 1,024 features, 1,083 features, 678 features, and 1,571 features, respectively. MCA was applied with the initial feature matrices to extract network-based features of drugs with at least 90% data variance preserved. As a result, the numbers of features extracted from the CSN, TPN, PWN, and DIN were 115, 86, 202, and 30 features, respectively.

3.2 Suitable numbers of hidden layer sizes

Based on information from each drug network, we developed an MLP model to classify whether a drug could be used to treat CRC or not. One parameter of the MLP model that should be optimally tuned to fit the data and directly affects the model performance is the number of hidden layer sizes, or the number of hidden nodes. To find the most suitable number of hidden layer sizes for each

drug-network-based model, we varied the number of hidden layer sizes as 10, 20, 30, 40, 50, and investigated the model performance by conducting 10-fold cross validation. The average AUC values of each model based on the different values of hidden layer sizes are shown in Figure 4.

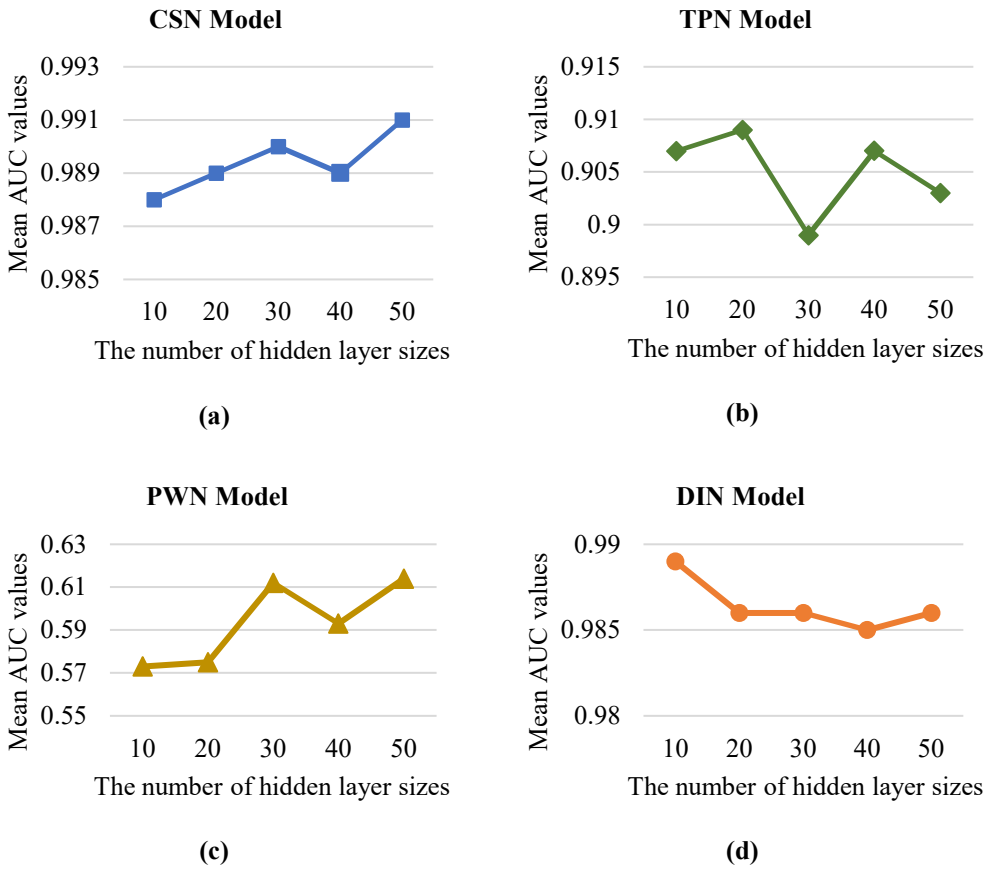

Figure 4. The mean AUC values of each model based on the different number of hidden layer sizes

According to Figure 4(a), the most suitable number of hidden layer sizes for the CSN model was 50, which gave the mean AUC value of 0.991. For the TPN model (Figure 4(b)), the highest mean AUC value of 0.909 was for the model that had 20 hidden nodes. In Figure 4(c), the PWN model reached the greatest mean AUC value of 0.614 when the number of hidden layer sizes was 50. In the case of the DIN model (Figure 4(d)), we obtained the maximum average AUC value of 0.989 when we used only 10 hidden nodes in the MLP model. Therefore, the selected numbers of hidden layer sizes that we used to develop the CSN, TPN, PWN, and DIN model were 50, 20, 50, and 10, respectively.

3.3 Performance of drug-network-based models

With the suitable number of hidden layer sizes, we constructed each drug-network-based model and evaluated its performance by using 10-fold cross validation. The percentages of the average

performance values were compared among four drug-network-based models, as shown in Table 2, to find the best model for further identifying potentially repurposable drug candidates for CRC. Note that bold percentages in Table 2 are the superior performance values when compared among all four models.

Table 2. The averages and standard deviations (SD) of performance values of each drug-network-based model

Performance Metrics	Model Names							
	CSN		TPN		PWN		DIN	
	Average (%)	SD (%)	Average (%)	SD (%)	Average (%)	SD (%)	Average (%)	SD (%)
AUPR	**99.3**	0.4	87.6	4.3	65.3	5.4	99.1	1.3
AUC	**99.1**	0.7	90.9	3.0	61.4	7.0	98.9	1.7
Precision	93.6	4.3	87.2	3.4	67.0	3.2	**96.5**	3.8
Recall	97.8	1.4	95.9	4.0	**99.1**	1.4	97.5	2.3
Accuracy	95.5	2.5	90.8	2.1	74.9	3.8	**96.9**	2.2
MCC	91.2	4.9	82.3	4.1	57.0	5.9	**93.9**	4.3
F_1 score	95.6	2.4	91.3	2.0	79.9	2.2	**96.9**	2.1

A bold and underlined number indicates the maximum value of a performance metric.

From Table 2, both the CSN and DIN models noticeably outperformed the others, with both producing average performance values above 90% for all evaluation metrics. The CSN model achieved the maximum AUPR value of 99.3% and the maximum AUC value of 99.1%, whereas the DIN model reached slightly lower AUPR and AUC values of 99.1% and 98.9%, respectively. The chemical structures of drugs affect their physicochemical properties and pharmacological activities [25]. This is the reason why drug chemical structures are usually used to infer drug targets, indications, side effects, drug interactions, and drug repurposability. However, by performing paired *t*-tests, we found that both mean AUPR and AUC values of the CSN and DIN model were not significantly different.

When the values of other evaluation metrics were compared, it was noticeable that the DIN model outperformed the others with a precision of 96.5%, a recall of 97.5%, an accuracy of 96.9%, an MCC of 93.9%, and an F_1 of 96.9%. Because drug-drug interactions are molecular data possibly inferring to physiological effects and drug targets [26], the model based on DIN information could effectively be used to classify repurposable drug candidates for CRC. Although the PWN model achieved the greatest recall value of 99.1%, the PWN model yielded the lowest values of other evaluation metrics, such as a precision value of only 67.0% and an accuracy value of 74.9%. This could be because the PWN model produced many false positives as it predicted too many repurposable drugs for CRC, resulting in a high recall value but low precision and accuracy values. Therefore, we selected the DIN model as the best-performing model that could be used for the prediction of repurposable drug candidates for CRC.

3.4 Comparison with existing methods for generic drug repurposing

In this experiment, the performance of the DIN model was compared with those of some existing methods created for generic purposes in drug repurposing. The existing models were a three-layer heterogeneous network model (TL-HGBI) [9], a similarity network model with bi-random walks (MBiRW) [10], and a method of topological similarity and singular value decomposition (TS-SVD) [11]. For training and testing the models, we used the data set of Kawichai *et al*. [12], which combined drug-disease associations of numerous diseases from more than one source. For the testing data set predictions, only predictions involved with CRC were included in computing the performance values of TL-HGBI, MBiRW, and TS-SVD models. The averages and standard deviations (SD) of performance values of the TL-HGBI, MBiRW, TS-SVD, and DIN models are shown in Table 3.

Table 3. The averages and standard deviations (SD) of performance values of three existing models for generic drug repurposing and the DIN model

Performance Metrics	Model Names							
	TL-HGBI		MBiRW		TS-SVD		DIN	
	Average (%)	SD (%)	Average (%)	SD (%)	Average (%)	SD (%)	Average (%)	SD (%)
AUPR	89.3	3.7	97.2	1.4	98.1	1.3	**99.1**	1.3
AUC	77.0	5.3	92.8	3.1	94.0	3.6	**98.9**	1.7
Precision	85.2	3.5	93.2	1.8	94.1	2.2	**96.5**	3.8
Recall	97.2	2.2	**99.3**	1.2	97.8	1.8	97.5	2.3
Accuracy	85.3	2.7	94.0	2.0	93.5	2.1	**96.9**	2.2
MCC	52.4	12.1	82.2	6.4	78.8	6.2	**93.9**	4.3
F_1 score	90.6	1.8	96.1	1.3	95.8	1.4	**96.9**	2.1

A bold and underlined number indicates the maximum value of a performance metric.

According to Table 3, it is noticeable that the DIN model outperformed the others with the highest values of most performance metrics except for recall. The results indicate that the DIN model, which was particularly developed for CRC drug repurposing, is more effective in uncovering repurposable drugs for CRC than the existing methods (i.e., TL-HGBI, MBiRW, and TS-SVD) which were developed for generic purposes in drug repurposing.

3.5 Repurposable drug candidates for colorectal cancer

To identify repurposable drug candidates for CRC, we rebuilt the DIN model with 22 CRC approved drugs and 321 non-CRC approved drugs. Then, the model was employed to predict the remaining non-CRC approved drugs that were not used to train the DIN model and for which drug-drug interaction information was available. In total, there were 1,213 non-CRC approved drugs that were screened for the drug repurposability for CRC. As a result, 306 non-CRC approved drugs (25.2%) were predicted to be potential repurposable drugs for the treatment of CRC. The histogram based on the predicted scores of the CRC repurposable drug candidates is shown in Figure 5.

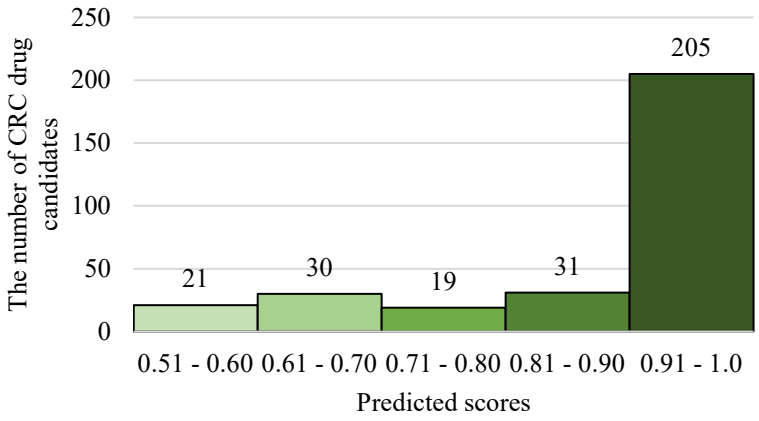

Figure 5. The histogram of predicted scores of CRC repurposable drug candidates

In Figure 5, the majority of CRC drug candidates (205 out of 306 or 67%) had predicted scores more than 0.9. From this group of the drug candidates, we selected four cases to demonstrate the validity of the predictions and the practicality of our drug-network-based models for identification of CRC repurposable drug candidates. The details of the selected cases and their supporting information are shown in Table 4. In the case of supporting information from ClinicalTrials.gov, we also provide the National Clinical Trial (NCT) numbers of related clinical studies in the Table for conveniently searching on more information.

Table 4. Four selected cases of the repurposable drug candidates for CRC

Drug Name (DrugBank ID)	FDA-Approved Indications	Supporting Information
Lurbinectedin (DB12674)	Metastatic small-cell lung cancer (SCLC)	- Clinical studies in combination with capecitabine (NCT02210364) and irinotecan (NCT02611024) - Inhibition of DNA transcription, leading to apoptosis of tumor cells [27]
Talimogene laherparepvec (DB13896)	Recurrent melanoma	- Clinical studies in combination with atezolizumab (NCT03256344) and in patients with cutaneous squamous cell cancer (NCT03714828)
Nalmefene (DB06230)	Alcohol dependence	- Suppression of the glycolysis of CRC cells [28]
Verteporfin (DB00460)	Macular degeneration and presumed ocular histoplasmosis syndrome	- Suppression of YAP expression [29] - Activation of ROS and the JNK pathway [30]

Lurbinectedin (DB12674) is a DNA alkylating agent approved for the treatment of metastatic small-cell lung cancer and was investigated by researchers for the treatment of various types of cancer, such as breast cancer, pancreatic cancer, and CRC. In a clinical study with the ClinicalTrials.gov ID of NCT02210364, the pharmacokinetics of lurbinectedin were investigated in combination with capecitabine for the treatment of CRC. Additionally, the use of lurbinectedin with irinotecan was studied for the treatment of several advanced solid tumors, including CRC, in a

clinical study with the NCT number of NCT02611024. From *in vivo* studies using xenograft models, lurbinectedin showed strong apoptosis induction in multiple cancer cells, including CRC cells, by inhibiting DNA transcription and inducing DNA double-strand breaks [27].

Talimogene laherparepvec or T-VEC (DB13896) was the first viral immunotherapy approved for the treatment of recurrent melanoma. To support its suggested repurposability for CRC, we found a clinical study (NCT03256344) which evaluated the safety of injecting T-VEC in combination with atezolizumab, a medication for non-small cell lung cancer, in patients with metastatic CRC. Furthermore, there was another clinical study (NCT03714828) that investigated the use of T-VEC to treat patients with cutaneous squamous cell cancer, including hereditary nonpolyposis CRC.

Nalmefene (DB06230) is an opioid receptor antagonist for the treatment of patients with alcohol dependence. Although there was no clinical study involved with nalmefene in ClinicalTrials.gov, there were some studies concerning the repurposability of this drug. With the transwell migration assay, it was revealed that nalmefene can mitigate the growth of CT26 colon cancer cells by suppressing the expression of calmodulin and the phosphorylation of Ca^{2+}-calmodulin stimulated protein kinase II (CaMKII), leading to inhibition of glycolysis in CRC cells [28].

Verteporfin (DB00460) is a medication approved for the treatment of certain eye diseases, including macular degeneration, and presumed ocular histoplasmosis syndrome. There were some studies about the effects and mechanisms of verteporfin in destroying tumor cells. For example, Takeda *et al*. [29] demonstrated that verteporfin suppressed the expression of yes-associated protein (YAP) resulting in inhibition of migration and invasion of a CRC cell line DLD-1. Song *et al*. [30] revealed that the use of verteporfin as a photosensitizing agent in a photodynamic therapy could inhibit the proliferation of CRC cells *in vitro* and *in vivo* by inducing significant generation of reactive oxygen species (ROS) in CRC cells through regulating the c-Jun N-terminal kinase (JNK) and mechanistic target of rapamycin (mTOR) pathway.

Furthermore, we searched for supporting evidence of 306 repurposable drug candidates for CRC from ClinicalTrials.gov and published literature to bolster the model's validity. We found that 160 out of 306 drug candidates (52.3%) had supporting evidence in ClinicalTrials.gov, and 193 out of 306 drug candidates (63.1%) had supporting literature. Additionally, we found that 138 out of 306 drug candidates (45.1%) had supporting evidence in both ClinicalTrials.gov and published literature.

4. Conclusions

In this work, four classification models were developed based on four drug networks. The models were the CSN, TPN, PWN, and DIN, and the purpose of these models was the discovery of repurposable drug candidates for CRC. The results showed that the DIN model outperformed the others, having the highest accuracy and F_1 score of 96.9%. After classifying over 1,200 non-CRC approved drugs with the DIN model, we discovered 306 potentially repurposable drug candidates for CRC. From this group, lurbinectedin and talimogene laherparepvec (T-VEC) had been investigated in clinical studies for the treatment of CRC. In addition, two compounds, nalmefene and verteporfin were reported for their underlying mechanisms in suppressing the growth of CRC cells. In summary, the drug-network-based classification models can be used to effectively classify CRC repurposable drugs and can be applied to conduct large-scale screening of compounds, not only limited for small molecule compounds, but also for large biological compounds. The future direction of this work includes developing positive-unlabeled classification models to produce more accurate predictions of repurposable drugs for CRC and other types of cancer.

References

[1] Marmol, I., Sanchez-de-Diego, C., Pradilla Dieste, A., Cerrada, E. and Rodriguez Yoldi, M.J., 2017. Colorectal carcinoma: A general overview and future perspectives in colorectal cancer. *International Journal of Molecular Sciences,* 18(1), https://doi.org/10.3390/ijms18010197.

[2] Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A. and Bray, F., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians,* 71(3), 209-249, https://doi.org/10.3322/caac.21660.

[3] Islam, M.R., Akash, S., Rahman, M.M., Nowrin, F.T., Akter, T., Shohag, S., Rauf, A., Aljohani, A.S.M. and Simal-Gandara, J., 2022. Colon cancer and colorectal cancer: Prevention and treatment by potential natural products. *Chemico-Biological Interactions,* 368, https://doi.org/10.1016/j.cbi.2022.110170.

[4] Morgan, E., Arnold, M., Gini, A., Lorenzoni, V., Cabasag, C.J., Laversanne, M., Vignat, J., Ferlay, J., Murphy, N. and Bray, F., 2023. Global burden of colorectal cancer in 2020 and 2040: incidence and mortality estimates from GLOBOCAN. *Gut,* 72(2), 338-344, https://doi.org/10.1136/gutjnl-2022-327736.

[5] Juneja, M., Kobelt, D., Walther, W., Voss, C., Smith, J., Specker, E., Neuenschwander, M., Gohlke, B.O., Dahlmann, M., Radetzki, S., Preissner, R., von Kries, J.P., Schlag, P.M., and Stein, U., 2017. Statin and rottlerin small-molecule inhibitors restrict colon cancer progression and metastasis via MACC1. *PLOS Biology,* 15(6), https://doi.org/10.1371/journal.pbio.2000784.

[6] Hecht, M., Harrer, T., Buttner, M., Schwegler, M., Erber, S., Fietkau, R. and Distel, L.V., 2013. Cytotoxic effect of efavirenz is selective against cancer cells and associated with the cannabinoid system. *AIDS,* 27(13), 2031-2040, https://doi.org/10.1097/qad.0b013e3283625444.

[7] Tao, C., Sun, J., Zheng, W.J., Chen, J. and Xu, H., 2015. Colorectal cancer drug target prediction using ontology-based inference and network analysis. *Database,* 2015, https://doi.org/10.1093/database/bav015.

[8] Irham, L.M., Wong, H.S., Chou, W.H., Adikusuma, W., Mugiyanto, E., Huang, W.C. and Chang, W.C., 2020. Integration of genetic variants and gene network for drug repurposing in colorectal cancer. *Pharmacological Research,* 161, https://doi.org/10.1016/j.phrs.2020.105203.

[9] Wang, W., Yang, S., Zhang, X. and Li, J., 2014. Drug repositioning by integrating target information through a heterogeneous network model. *Bioinformatics,* 30(20), 2923-2930, https://doi.org/10.1093/bioinformatics/btu403.

[10] Luo, H., Wang, J., Li, M., Luo, J., Peng, X., Wu, F.-X. and Pan, Y., 2016. Drug repositioning based on comprehensive similarity measures and bi-random walk algorithm. *Bioinformatics,* 32(17), 2664-2671, https://doi.org/10.1093/bioinformatics/btw228.

[11] Liu, J., Zuo, Z. and Wu, G., 2020. Link prediction only with interaction data and its application on drug repositioning. *IEEE Transactions on NanoBioscience,* 19(3), 547-555, https://doi.org/10.1109/TNB.2020.2990291.

[12] Kawichai, T., Suratanee, A. and Plaimas, K., 2021. Meta-path based gene ontology profiles for predicting drug-disease associations. *IEEE Access,* 9, 41809-41820, https://doi.org/10.1109/ACCESS.2021.3065280.

[13] Yi, H.-C., You, Z.-H., Wang, L., Su, X.-R., Zhou, X. and Jiang, T.-H., 2021. In silico drug repositioning using deep learning and comprehensive similarity measures. *BMC Bioinformatics,* 22(3), https://doi.org/10.1186/s12859-020-03882-y.

[14] Zhao, B.-W., Su, X.-R., Hu, P.-W., Ma, Y.-P., Zhou, X. and Hu, L., 2022. A geometric deep learning framework for drug repositioning over heterogeneous information networks. *Briefings in Bioinformatics,* 23(6), https://doi.org/10.1093/bib/bbac384.

[15] Wishart, D.S., Feunang, Y.D., Guo, A.C., Lo, E.J., Marcu, A., Grant, J.R., Sajed, T., Johnson, D., Li, C., Sayeeda, Z., Assempour, N., Iynkkaran, I., Liu, Y., Maciejewski, A., Gale, N., Wilson, A., Chin, L., Cummings, R., Le, D., Pon, A., Knox, C. and Wilson, M., 2018.

DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research,* 46(D1), D1074-D1082, https://doi.org/10.1093/nar/gkx1037.

[16] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B.A., Thiessen, P.A., Yu, B., Zaslavsky, L., Zhang, J. and Bolton, E.E., 2023. PubChem 2023 update. *Nucleic Acids Research,* 51(D1), D1373-D1380, https://doi.org/10.1093/nar/gkac956.

[17] Gaulton, A., Hersey, A., Nowotka, M., Bento, A.P., Chambers, J., Mendez, D., Mutowo, P., Atkinson, F., Bellis, L.J., Cibrian-Uhalte, E., Davies, M., Dedman, N., Karlsson, A., Magarinos, M.P., Overington, J.P., Papadatos, G., Smit, I. and Leach, A.R., 2017. The ChEMBL database in 2017. *Nucleic Acids Research,* 45(D1), D945-D954, https://doi.org/10.1093/nar/gkw1074.

[18] Kanehisa, M. and Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research,* 28(1), 27-30, https://doi.org/10.1093/nar/28.1.27.

[19] Wishart, D.S., Li, C., Marcu, A., Badran, H., Pon, A., Budinski, Z., Patron, J., Lipton, D., Cao, X., Oler, E., Li, K., Paccoud, M., Hong, C., Guo, A.C., Chan, C., Wei, W. and Ramirez-Gaona, M., 2020. PathBank: a comprehensive pathway database for model organisms. *Nucleic Acids Research,* 48(D1), D470-D478, https://doi.org/10.1093/nar/gkz861.

[20] Xiong, G., Yang, Z., Yi, J., Wang, N., Wang, L., Zhu, H., Wu, C., Lu, A., Chen, X., Liu, S., Hou, T. and Cao, D., 2022. DDInter: an online drug-drug interaction database towards improving clinical decision-making and patient safety. *Nucleic Acids Research,* 50(D1), D1200-D1207, https://doi.org/10.1093/nar/gkab880.

[21] Newman, M., 2010. *Networks: An Introduction*. New York: Oxford University Press Inc.

[22] Steen, M.V., 2010. *Graph Theory and Complex Networks: An Introduction*. Enschede: Maarten Van Steen.

[23] Mori, Y., Kuroda, M. and Makino, N., 2016. Multiple correspondence analysis. In: N. Kunitomo and A. Takemura, eds. *Nonlinear Principal Component Analysis and Its Applications*. Singapore: Springer Singapore, pp. 21-28.

[24] Chawla, N.V., Bowyer, K.W., Hall, L.O. and Kegelmeyer, W.P., 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research,* 16(1), 321-357.

[25] Mao, F., Ni, W., Xu, X., Wang, H., Wang, J., Ji, M. and Li, J., 2016. Chemical structure-related drug-like criteria of global approved drugs. *Molecules,* 21(1), https://doi.org/10.3390/molecules21010075.

[26] Zhou, B., Wang, R., Wu, P. and Kong, D.X., 2015. Drug repurposing based on drug-drug interaction. *Chemical Biology and Drug Design,* 85(2), 137-144, https://doi.org/10.1111/cbdd.12378.

[27] Musacchio, L., Cicala, C.M., Salutari, V., Camarda, F., Carbone, M.V., Ghizzoni, V., Giudice, E., Nero, C., Perri, M.T., Ricci, C., Tronconi, F., Scambia, G. and Lorusso, D., 2022. Preclinical and clinical evidence of lurbinectedin in ovarian cancer: current status and future perspectives. *Frontiers in Oncology,* 12, https://doi.org/10.3389/fonc.2022.831612.

[28] Wu, Q., Chen, X., Wang, J., Sun, P., Weng, M., Chen, W., Sun, Z., Zhu, M. and Miao, C., 2018. Nalmefene attenuates malignant potential in colorectal cancer cell via inhibition of opioid receptor. *Acta Biochimica et Biophysica Sinica,* 50(2), 156-163, https://doi.org/10.1093/abbs/gmx131.

[29] Takeda, T., Yamamoto, Y., Tsubaki, M., Matsuda, T., Kimura, A., Shimo, N. and Nishida, S., 2022. PI3K/Akt/YAP signaling promotes migration and invasion of DLD-1 colorectal cancer cells. *Oncology Letters* 23(4), https://doi.org/10.3892/ol.2022.13226.

[30] Song, C., Xu, W., Wu, H., Wang, X., Gong, Q., Liu, C., Liu, J. and Zhou, L., 2020. Photodynamic therapy induces autophagy-mediated cell death in human colorectal cancer cells via activation of the ROS/JNK signaling pathway. *Cell Death and Disease,* 11(10), https://doi.org/10.1038/s41419-020-03136-y.