OptNCMiner: A deep learning approach for the discovery of natural compounds modulating disease-specific multi-targets

Seo Hyun Shin
Seoul National University
Seung Man Oh
Seoul National University
Jung Han Yoon Park
Seoul National University
Ki Won Lee
Seoul National University
Hee Yang (yhee6106@snu.ac.kr)
Seoul National University

Research Article

Keywords: Natural compounds, Chemical-protein interaction, Multi-target prediction, Deep learning, Siamese neural network

Posted Date: March 25th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1467387/v1

Abstract

Background: Due to their diverse bioactivity, natural product (NP)s have been developed as commercial products in the pharmaceutical, food and cosmetic sectors as natural compound (NC)s and in the form of extracts. Following administration, NCs typically interact with multiple target proteins to elicit their effects. Various machine learning models have been developed to predict multi-target modulating NCs with desired physiological effects. However, due to deficiencies with existing chemical-protein interaction datasets, which are mostly single-labeled and limited, the existing models struggle to predict new chemical-protein interactions. New techniques are needed to overcome these limitations.

Results: We propose a novel NC discovery model called OptNCMiner that offers various advantages. The model is trained via end-to-end learning with a feature extraction step implemented, and it predicts multi-target modulating NCs through multi-label learning. In addition, it offers a few-shot learning approach to predict NC-protein interactions using a small training dataset. OptNCMiner achieved better prediction performance in terms of recall than conventional classification models. It was tested for the prediction of NC-protein interactions using small datasets and for a use case scenario to identify multi-target modulating NCs for type 2 diabetes mellitus complications.

Conclusions: OptNCMiner identifies NCs that modulate multiple target proteins, which facilitates the discovery and the understanding of biological activity of novel NCs with desirable health benefits.

Background

Natural product (NP)s, are defined as substances produced by living organisms. The term NP is often used broadly, covering both natural compound (NC)s and mixtures of NCs derived from natural sources (1). The diverse biological activities of NPs are due to the activity of their constituent NCs. In order to fully understand and utilize NPs, it is important to identify and investigate the mode of action of active NCs.

Subject to selective pressure over millions of years, NCs now have diverse bioactive chemotypes, some of which are optimized for particular biological functions, such as endogenous growth and the defense of living organisms (2). Due to their wide range of target space and bioactivity, NCs have been an important foundation for traditional medicines and modern pharmacology. The World Health Organization has reported that approximately 20,000 plants are used for herbal remedies in 91 countries worldwide (3). Diverse bioactive scaffolds of NCs have been examined as drug candidates for the treatment of various diseases, including cancer, cardiovascular diseases, and infectious diseases (2). From 1981 to 2014, more than a half of all approved novel drugs were derived from NCs (4). In addition to pharmaceutical sources, NC applications are also expanding in the functional food, cosmetics and agricultural industries (5–8).

Another interesting characteristic of NCs is their propensity to modulate multiple protein targets. The unique scaffolds and structural motifs of NCs enable interactions with multiple target proteins to elicit diverse biological activities (9). In the past few decades, the paradigm of drug discovery has shifted from

'one target, one disease' to 'multi-target' or 'multimodal' drug discovery, especially for diseases with complex etiologies or drug resistance issues (10). Such multi-target modulating drugs have been expected to improve safety issues and enhance clinical efficacy compared to single-target drugs (11–13). For example, resveratrol, a well-known stilbenoid found in red wine and various foods, has been associated with 21 direct molecular targets including SIRT1 (14). Resveratrol has been shown in clinical studies to have beneficial effects on pathways implicated in a variety of diseases, such as diabetes, obesity, various types of cancer, Alzheimer's disease and cardiovascular disease (15). However, due to the NCs targeting multiple proteins, it has been challenging to discover optimal NCs, which regulates desired targets and avoids any off-targets

Various experimental methods have been developed over the past few decades to identify NCs that regulate target proteins (2). However, the discovery process for multi-target modulating NCs is tedious and cost-intensive. For translation of discoveries into drug development, a pre-approval process of lead identification, compound optimization, *in vitro* and animal experiments is required before the first clinical trial can be conducted. Developing a new drug through this process typically costs in excess of 800 million US dollars (16). In recent years, computational methods have provided substantial assistance in the discovery of new NCs. Molecular descriptors and fingerprint methods have made it possible to describe NCs in mathematical expressions (17), while 3D modeling and docking methods have been developed to simulate a complex molecular structure and conformational space of NCs, as well as their interactions with target proteins (18). These cheminformatics tools, along with chemical-protein interaction databases, have shed light on the development of machine learning models that predict novel NCs and streamline the NC discovery process.

A variety of machine learning methods have been trained on chemical-protein interaction databases to predict compounds with novel target modulating activities (19–21). Of particular note, deep neural networks (DNN)s have been widely applied in the field of active compound discovery, as they enable the automation of the feature engineering process that often becomes a bottleneck in conventional machine learning methods. DNN's capabilities with end-to-end learning – a process of training parameters jointly – enables the automation of complicated predictive procedures ranging from data pre-processing to the prediction of processed data. Thus, DNN efficiently generates or extracts important hidden features from the input vectors of the compounds responsible for their activities, circumventing the need for manually engineering the input features (22). However, conventional machine learning models including DNN still carry limitation: their performance greatly relies on the data quantity and quality (23, 24). Unfortunately, established chemical-protein interaction databases are biased towards the most well-studied proteins and compounds. Thus, not all proteins have sufficient data to reliably train machine learning models. In addition, most existing databases are optimized for binary classification methods as they only provide single or limited target protein data for each compound. Datasets from such databases treat unknown chemical-protein interactions as negative data. Consequently, predictive models trained on single-label data run the risk of predicting false-negative interactions when the interactions can be experimentally tested positive. Thus, the prediction of such interactions should be treated as a multi-label classification task. Considering that the interaction data provided by the existing databases is single-labeled and

limited in size, a model capable of learning multi-classification from single labeled data is required to overcome the current limitations of NC discovery.

Siamese neural network (SNN), first suggested in the early 1990s by Bromley and LeCun, is comprised of two identical networks, with one called ‘heads’, that accepts distinct input pairs and is an activation function called 'body', that concatenates the two heads (25). SNN is a powerful tool for two reasons: it enables similarity comparisons of complex data and can be applied to one-shot and few-shot learning. When comparing the similarity between two high-dimensional data points, the SNN learns the hidden representations of the two input vectors in a parallel fashion and compares the outputs at the end using a similarity metric such as cosine distance. Unlike models that use classification loss functions to classify between classes, SNNs use contrastive loss functions to learn to distinguish between inputs. Since SNN uses pairs of similar and dissimilar data points for training, it can achieve greater performance with fewer data points. In addition, since the model is trained to predict similarity between input pairs, the performance of the model is not impacted by the class imbalance in positive and negative data. Due to these characteristics of SNN, it is a significantly compatible model, especially for learning protein-interacting compound data. SNN has been applied to various fields including image analysis, audio and speech processing, and sensor-based activity recognition (26–29). Furthermore, SNN has also been recently applied in the field of pharmacology. ReSimNet, a model for drug discovery and repositioning was developed by Jeon et al. in 2019 (30). Jeon and colleagues used SNN to predict transcriptional response similarities between two compounds using gene expression data from the CMap database. However, there has not yet been a case where an SNN has been applied to predict NCs that modulate multiple disease-specific target proteins.

Here, we describe ‘OptNCMiner’, a machine learning model suitable for predicting ‘optimal NCs’ that modulate disease-specific multi-targets. Built on a structure of SNN, OptNCMiner preserves the advantages of DNN to effectively extract essential features of NCs related to chemical-protein interactions. OptNCMiner is validated on its ability with multi-label learning from single positive data on chemical-protein interactions, and is also capable of few-shot learning, enabling multi-class classification on NCs using small datasets. We tested OptNCMiner with the discovery of natural sources containing NCs that regulate target proteins associated with type 2 diabetes mellitus (T2DM)-related complications.

Methods

OptNCMiner learns structural similarities between compound pairs and is trained to grant high similarity scores to chemical pairs where both compounds are active against a single target protein. Upon successful training, OptNCMiner calculates the activity score of NCs with each target protein in of the context of a similarity score between NCs and known active compounds of target proteins (Fig. 1). OptNCMiner was trained with three datasets of different sizes in order to test its learning capability regardless of dataset size. Comparisons with traditional classification models and validation of false positives using *in silico* docking simulation revealed that OptNCMiner successfully predicts both known and unknown chemical-protein interactions.

Data collection and preparation

Chemical-protein interaction data was gathered from ExCAPE-DB and LIT-PCBA. ExCAPE-DB is a database of chemogenomics data curated from two major public databases: PubChem and ChEMBL (31). Compound data curated from ExCAPE-DB were labeled according to the activity flag provided by ExCAPE-DB: ‘A’ for active and ‘N’ for inactive. LIT-PCBA is a dataset designed for virtual screening and machine learning based on PubChem bioassay data (32). Similarly, compound data curated from LIT-PCBA were labeled as active or inactive based on the given activity class for each compound.

Three datasets of different sizes were prepared and named as the 'base dataset', 'transfer learning dataset’, and 'few-shot learning dataset’ in accordance with their decreasing size (Table 1). All datasets were composed of active and inactive compounds. The base dataset was constructed with chemical-protein interaction data for 11 proteins and data for more than 5,000 active compounds, randomly selected from ExCAPE-DB. The transfer learning dataset was constructed with chemical-protein interaction data for 7 proteins with the number of actives between 500 and 1,000. The transfer learning dataset was used for transfer learning of OptNCMiner and three baseline models capable of multi-class classification: Cosine similarity, Random Forest (RF), and Multi-layer Perceptron (MLP). Data for the transfer learning dataset were collected from ExCAPE-DB and LIT-PCBA. The performance of OptNCMiner was compared to that of the baseline models. The few-shot learning dataset was constructed with chemical-protein interaction data for 7 proteins with data for less than 100 active compounds. The few-shot learning dataset was used for the few-shot learning of OptNCMiner and was collected from LIT-PCBA data. This set consists of 296 compounds in total.

Table 1
Data used to construct the base dataset, transfer learning dataset, and few-shot learning dataset.

Dataset	Target Gene	Target Protein	Data Source	Active compounds	Inactive compounds
Base dataset (actives > 5,000)	ADORA2A	Adenosine receptor A2a	ExCAPE-DB	5077	591
	BRCA1	Breast cancer type 1 susceptibility protein	ExCAPE-DB	8619	43095
	CNR1	Cannabinoid receptor 1	ExCAPE-DB	5125	397
	DRD2	D(2) dopamine receptor	ExCAPE-DB	8037	40185
	HTR1A	5-hydroxytryptamine receptor 1A	ExCAPE-DB	6339	31695
	KCNH2	Potassium voltage-gated channel subfamily H member 2	ExCAPE-DB	5327	26635
	LMNA	Prelamin-A/C	ExCAPE-DB	14533	72665
	OPRM1	Mu-type opioid receptor	ExCAPE-DB	5665	2872
	SLC6A4	Sodium-dependent serotonin transporter	ExCAPE-DB	6912	370
	TARDBP	TAR DNA-binding protein 43	ExCAPE-DB	12193	60965
	TDP1	Tyrosyl-DNA phosphodiesterase 1	ExCAPE-DB	23129	115645
Transfer learning dataset (1,000 > actives > 500)	ADRA2A	Alpha-2A adrenergic receptor	ExCAPE-DB	816	39
	GRIN1	Glutamate receptor ionotropic	ExCAPE-DB	553	92
	HTR3A	5-hydroxytryptamine receptor 3A	ExCAPE-DB	565	65
	MINK1	Misshapen-like kinase 1	ExCAPE-DB	929	8
	PKM2	Pyruvate kinase PKM	ExCAPE-DB	546	2730
	POLK	DNA polymerase kappa	LIT-PCBA	772	3860

Dataset	Target Gene	Target Protein	Data Source	Active compounds	Inactive compounds
	VDR	Vitamin D3 receptor	LIT-PCBA	884	4420
Few-shot learning dataset (100 > actives)	ADRB2	Beta 2 adrenergic receptor	LIT-PCBA	17	170
	ESR	Estrogen receptor alpha	LIT-PCBA	13	130
	IDH1	Isocitrate dehydrogenase	LIT-PCBA	39	390
	MTOR	mammalian target of rapamycin complex 1	LIT-PCBA	97	970
	OPRK1	Kappa opioid receptor	LIT-PCBA	24	5460
	PPARG	Peroxisome proliferator-activated receptor gamma	LIT-PCBA	27	270
	TP53	Cellular tumor antigen p53	LIT-PCBA	79	790

Input generation

For each compound, a standard fingerprint of 1024 bits was generated from the canonical SMILES representation using Chemistry Development Kit (CDK) in R software (33). 90% of compound pairs from the base dataset and transfer learning dataset served as the training set to construct the predictive model. The remaining 10% were used as an external validation set to evaluate the capability of the predictive model.

We have generated pairs of compounds and their labels, since OptNCMiner is a network that accepts inputs in the form of pairs and computes the similarity between the two. Compound pairs were labeled as ‘positive’ if both were classified as active for the same target protein. Those pairs that did not satisfy the criteria were labeled as ‘negative’. In order to prevent proteins with large interaction data sizes from dominating the training dataset, proteins were randomly sampled from a uniform distribution. For generating positive pairs, active compounds to the protein, which is shown as C_p in Fig. 1, were randomly sampled. On the other hand, compounds interacting with different proteins or negative compounds to the target protein (C_N in Fig. 1) were randomly chosen to generate negative pairs. 7,000 positive and negative pairs each were generated for the training dataset, in the form of fingerprint vectors of compound pairs concatenated with the binary labels of either 1 or 0.

Model building

OptNCMiner is built in an SNN structure, where pairs of inputs are fed to identical multi-layer perceptrons called ‘head function’ to generate pairs of embedding vectors. The similarity between two embedding

vectors is computed by a distance function referred to as the ‘body function’. The overall structure of the model is represented as follows:

$Y = B(H(X_1), H(X_2))$ [1]

where X_1, X_2 is a pair of chemical inputs in the form of fingerprint vectors, Y is the binary label of the pair, H(·) is the head function, and B(·) is the body function. The head function H(·) maps input vectors X_1, $X_2 \in R^{2048}$ into embedding vectors Z_1, $Z_2 \in R^{2048}$. The hidden layer dimension for H(·) is set to 2048-2048-2048 with a dropout of 0.5, and was constructed with PyTorch. The resulting embedding vectors are created in forms of $Z_1 = W_2 f(W_1 X_1 + b_1 + b_2)$and $Z_2 = W_2 f(W_1 X_2 + b_1 + b_2)$, where f(·) is a ReLU activation function and $W_1 \in R^{h \times 2048}$, $W_2 \in R^{e \times h}$, $b_1 \in R^h$, $b_2 \in R^e$ are trainable weights and biases, respectively. $Z_1$and Z_2 are then fed in to the body function B(·), which is a function of cosine distance, defined by equation [2]. We then defined the structural similarity of two compounds as the cosine distance between two embedding vectors. Finally, sigmoid function with equation [3] is used to produce the output called 'similarity score', where 1 refers to the presence of chemical-protein interaction, and 0 refers to its absence.

cosine distance\left({Z}_{1},{Z}_{2}\right)=\frac{{Z}_{1}\bullet {Z}_{2}}{‖{Z}_{1}‖\times ‖{Z}_{2}‖}=\frac{\sum _{i=1}^{n}{{Z}_{1}}_{i}\times {{Z}_{2}}_{i}}{\sqrt{\sum _{i=1}^{n}{{{Z}_{1}}_{i}}^{2}}\times \sqrt{\sum _{i=1}^{n}{{{Z}_{2}}_{i}}^{2}}} [2]

{\sigma }\left(\text{x}\right)= \frac{1}{1+{e}^{-x}} [3]

Binary Cross Entropy (BCE) was used as the loss function [4]. Through the optimization process, the model was trained to minimize BCE between the predicted output and the label. The model hyperparameters were optimized using the validation set during training. We used Adam optimizer (34) with a learning rate of 0.0001.

\text{B}\text{C}\text{E}= -\frac{1}{N}\sum _{i=0}^{N}{y}_{i}\bullet \text{log}\left({\widehat{y}}_{i}\right)+(1-{y}_{i})\bullet \text{l}\text{o}\text{g}(1-{\widehat{y}}_{i}) [4]

The training data are utilized as a support set and query set for testing upon completion of training. A support set was constructed by randomly selecting 100 compounds interacting with each protein. For each compound in a query set, its similarity to support set is computed. Then, for each protein, the maximum similarity score of the query compound is used to classify whether the compound binds to the protein. A threshold of 0.5 is used to determine the binding of the compound to the target protein.

Training approaches for varying dataset sizes: Transfer learning and few-shot learning

For many machine learning problems, better performance can be achieved by applying a transfer learning method, which pre-trains the model on a larger dataset before further training it on the target dataset (35).

We previously constructed the transfer learning dataset, which is composed of chemical-protein interaction data smaller in size (500 ~ 1000 active compounds) than the base dataset. We have adopted the idea of transfer learning on the transfer learning dataset, by pre-training the model with the base datasets and fine-tuning it with the transfer learning dataset.

For datasets too small to feasibly train on, few-shot learning can be used. We used the model trained for transfer learning to predict chemical-protein interactions with the few-shot learning dataset, taking the sparse available data and using it as a support set. 90% of the available data for the few-shot learning dataset was used as support and was tested on the remaining 10%.

Multi-label classification from single label data

Chu and colleagues have provided an updated gold standard dataset of chemical-protein interaction data that accounts for multiple bindings (36). To test the viability of learning multi-label representation from single-label data, the performance of OptNCMiner learning from a filtered single-label version of this data tested on the single-label test data was compared to the testing performance of the full multi-label data. performance with both single-label and multi-label datasets is shown in Additional file 1: Table 1 and all datasets used for the trial are available at the GitHub address listed in 'Availability of data and materials' section.

Model evaluation metrics

Since OptNCMiner is a classification model, typical classification metrics have been used to evaluate the performance of the model. A recall is a metric that measures the proportion of true positives, {T}_{p}, against all existing positives [5]. Thus, recall is a metric that evaluates misclassification of actual positives. Accuracy measures the general performance of the model, by calculating the proportion that the model classifies correctly among the entirety of the predictions [6]. The area under the receiver operating characteristics (AUROC) is an area under the curve drawn on the plot between the true positive rate (TPR) and false positive rate (FPR). The AUROC value represents the degree of class separability of the model. The final evaluation metrics are calculated as the weighted average of metric values of all proteins.

\text{R}\text{e}\text{c}\text{a}\text{l}\text{l}= \frac{{T}_{p}}{{T}_{P}+{F}_{N}} [5]

\text{A}\text{c}\text{c}\text{u}\text{r}\text{a}\text{c}\text{y}= \frac{{T}_{p}+{F}_{P}}{{T}_{P}+{T}_{N}+{F}_{P}+{F}_{N}} [6]

Results And Discussion

Dataset analysis

The chemical-protein interaction data from three differently-sized datasets are aggregated into a total of 106,317 positive interactions with 25 different target proteins. Since all proteins have different instances

of interactions, 1,000 active compounds were randomly selected for each protein in the base dataset, and all active compound data from the transfer learning data and few-shot learning data were used to investigate the target protein-interacting compound space.

To explore the physicochemical properties of the sampled compounds, principle component analysis (PCA) was used. The following ten physicochemical properties were calculated using OPERA 2.6 (37) to generate the PCA plot: Octanol-water partition coefficient (LogP), melting point (MP), boiling point (BP), vapor pressure (VP), water solubility (WS), Henry's Law constant (HL), Octanol-air partition coefficient (KOA), retention time (RT), acid dissociation constant (pKa), and pH-dependent lipid-aqueous partition coefficient (LogD). The scattered compounds are represented (in Fig. 2a) as a 2D diagram of PCA, where the compounds are assigned with different color codes according to the proteins they interact with.

Additionally, the structural similarity distribution of the sampled compounds was investigated using a pairwise Tanimoto similarity matrix. Sampled compounds were represented in a high dimensional space with a 1,024-bit standard fingerprint, containing information for the chemical substructures. All possible pairs of compounds were generated from those sampled and Tanimoto similarity values between 1,024-bit standard fingerprint vectors of chemical pairs were calculated. The resultant similarity matrix was rendered in the form of a heat map (Fig. 2b), where the sampled compounds were positioned in an order of base dataset, a transfer learning dataset, and a few-shot learning dataset.

The PCA plot in Fig. 2(a) reveals that compounds interacting with 25 different proteins share similar physicochemical properties and cannot easily diverge. However, the Tanimoto similarity values calculated from the fingerprint vectors of the compound pairs are distributed mostly between 0.1 to 0.3, which means almost all of the sampled compounds are structurally different from each other. From Fig. 2(a) and 2(b), we can conclude that our compounds share similar physicochemical properties but are structurally diverse. Thus, there was no predetermined structures or physicochemical properties that facilitated the prediction of the different target proteins. Using such compound data, OptNCMiner was trained to learn these hidden features to discern the binding natures of the compounds with different target proteins.

Performance evaluation

To evaluate whether OptNCMiner is sufficiently specialized for NC multi-target prediction, we examined its ability to learn multi-label classification by comparing the performance of the model after training with single-label and multi-label data. The single-label data were generated by deleting classes – in this case, target proteins of compounds from the original multi-label data. In a manner similar to how Cole validated multi-label learning with single-label data, we show that OptNCMiner has the ability to identify multiple targets for compounds from single-label data (38). OptNCMiner achieved similar recall and AUROC values when trained on single-label and multi-label data and then tested with the multi-label test dataset (Additional file1: Table S1).

The performance of OptNCMiner was also evaluated with the compounds not used in the training pair generation from the base dataset and transfer dataset. Our framework allows for the prediction of multiple binding targets of NCs not included in the original binding data; since the test data is labeled in a single-positive manner, the predicted output contains a high number of false positives. Therefore, recall (the ratio of predicted positives among the entire positives) is considered to be the most relevant evaluation metric.

The performance of OptNCMiner was compared to three baseline models capable of multi-label classification, which are cosine similarity, RF, and MLP. The baseline models were trained with 90% and were evaluated using the remaining 10% of the base and transfer-learning datasets. For the cosine similarity method, the cosine similarity of two standard fingerprint vectors was calculated. If the cosine similarity value of a pair exceeded 0.5, two compounds were considered structurally similar and the test compound is predicted to bind to the target protein, and vice versa. RF and MLP were trained to classify compound pairs in binary labels, either 1 or 0. A binary label of 1 refers to the test chemical being similar to its pair and thus, predicted to bind to the same target protein. The performance of the baseline models was evaluated with the base dataset and transfer learning dataset. All three baseline models were first trained and evaluated with the base dataset, while RF and MLP were further trained in the manner of the transfer learning with the transfer learning dataset. Since the cosine similarity method is not capable of transfer learning, training and evaluation were performed with the transfer learning dataset.

Table 2 shows that OptNCMiner outperformed the baseline models. The recall value of OptNCMiner for the base dataset was 0.833 and 0.871 for the transfer learning dataset, which means over 80% of known positives were predicted correctly. The second best recall value was achieved by RF with 0.677 for the base dataset, and MLP by 0.642 for the transfer learning dataset. The values of AUROC for OptNCMiner on both the base dataset and transfer learning dataset are above 50%, showing the model's discriminatory powers to be better than random chance. The relatively low accuracy and AUROC values may be due to newly predicted compound targets, which result in high false positive rates. MLP and cosine similarity generated accuracy values over 0.7, but relatively low AUROC values, suggesting that both methods are vulnerable to data imbalances. All evaluation metric values have been improved in the transfer learning dataset compared to the base dataset for OptNCMiner. Considering that both the base dataset and the transfer learning dataset held diverse compound structures, the improved evaluation metrics denote that the performance of the model was improved by the transfer learning method, regardless of the data characteristics.

Another strong advantage of OptNCMiner is its ability to predict chemical-protein interactions for proteins with limited training data. The trained and transfer-learned OptNCMiner using the base dataset and transfer learning dataset was used to predict chemical-protein interactions in the few-shot learning dataset using a few-shot learning method. The few-shot learning performance of OptNCMiner was evaluated with 7 proteins with data for less than 100 interacting compounds and is demonstrated in Table 3. OptNCMiner achieved 0.829 for the weighted average of recall and 0.665 for the weighted AUROC average. Although Beta 2 adrenergic receptor and Isocitrate dehydrogenase show relatively poor

performance, the model's predictions are remarkably accurate for most proteins. The performance is not necessarily correlated to the number of available data points, which is shown in the last column of Table 3, indicating that either the presence of some interactions were not identified by the network, or is indicative of a lack of representation present in the small data. Thus, it is confirmed that OptNCMiner possesses the ability to identify structural properties of compounds that enable specific chemical-protein interactions from a small number of samples.

Table 2
The performance of OptNCMiner and baseline models with the Base Dataset and Transfer Learning Dataset

Model	Performance metric[1]	Base dataset	Transfer learning dataset
OptNCMiner	Recall	0.833	0.871
	AUROC	0.632	0.787
	Accuracy	0.440	0.713
Cosine Similarity	Recall	0.573	0.696
	AUROC	0.643	0.761
	Accuracy	0.708	0.818
Random Forest	Recall	0.677	0.479
	AUROC	0.343	0.241
	Accuracy	0.028	0.027
Multi-Layer Perceptron	Recall	0.361	0.642
	AUROC	0.676	0.817
	Accuracy	0.972	0.974

[1]All performance metrics are weighted averages of the results of all proteins comprising the dataset.

Table 3
Performance of OptNCMiner with the few-shot learning dataset

Target Protein	Recall	AUROC	Accuracy	Count
Beta 2 adrenergic receptor	0.488	0.400	0.450	5
Estrogen receptor a	0.585	1.000	0.764	5
Isocitrate dehydrogenase	0.488	0.600	0.536	5
Mammalian target of rapamycin complex 1	0.537	0.889	0.663	9
Kappa opioid receptor	0.659	1.000	0.806	5
Peroxisome proliferator-activated receptor gamma	0.610	1.000	0.778	5
Cellular tumor antigen p53	0.537	0.857	0.664	7
Weighted average	**0.555**	**0.829**	**0.665**	**41**

Output validation

OptNCMiner is a model capable of predicting multiple target proteins for NCs through multi-label learning. OptNCMiner predicts not only the known protein targets of NCs, but also unknown target proteins, which results in high false positive rates. To confirm OptNCMiner's ability to predict unknown chemical-protein interactions, we sought to validate OptNCMiner's false positive outputs in few-shot learning.

OptNCMiner went through few-shot learning using the few-shot learning dataset, where chemical-protein interactions were predicted among 7 proteins and 23 compounds in the test set. To examine whether the false positives were real negatives or newly discovered target proteins, literature searches and *in silico* docking using GalaxyDockWEB (39) were undertaken. Among the 115 false positives, 4 chemical-protein interactions were validated in the literature (Additional file 1: Table S2). All 115 compounds went through *in silico* docking. As shown in Fig. 3(a), 114 of the 115 chemical-protein interactions generated a negative binding affinity score (see Additional file 2), suggesting that binding of the ligand to the active site exists in a favorable energy state (40). The result of *in silico* docking indicates that the false positives predicted by OptNCMiner are real unknown positives with a high probability. Two examples of successful docking of compounds and target proteins, where chemical-protein interactions were previously unknown, are illustrated in Fig. 3(b). The examples were selected based on the two lowest binding affinities among the 115 false positives, which were − 29.307 and − 28.121 respectively. In both examples, the compounds DIPPDP and ACPPTN (dark grey), are stably docked in the pre-assigned binding pockets with interacting amino acid side chains of the target proteins, estrogen receptor alpha and beta 2 androgenic receptor (light green).

Use case scenario: NCs present in natural sources that modulate target proteins associated with T2DM complications

To investigate the practical application of OptNCMiner in novel NC discovery, the program was trained using the base dataset to identify NCs present in natural sources that modulate target proteins associated with T2DM complications. Diabetic nephropathy, diabetic keratopathy, and cardiomyopathy were chosen as T2DM complications and target proteins for each complication were identified based on previous reports (41). Among the identified target proteins, 8 proteins with interacting compound data were selected and used as target protein candidates: Peroxisome proliferator activated receptor α (PPARα), Yes-associated protein (YAP), Phosphoinositide 3-kinase (PI3K), Protein kinase C β (PKCβ), Toll-like receptor 4 (TLR4), Sodium/glucose cotransporter 2 (SGLT2), G protein-coupled receptor 120 (GPR120), and Nuclear factor erythroid 2-related factor 2 (Nrf-2) (Additional file 1: Table S3). The interacting compound data was gathered from BindingDB, a chemical-protein interaction database (42). The sizes of interacting compound data varied among the target proteins. Transfer learning or few-shot learning was applied according to the two different ranges of data size. Five target proteins with more than 100 interacting compound data points were assigned to transfer learning and three target proteins with less data were assigned to few-shot learning. In order to predict NCs in natural sources, data including NCs and natural resources containing NCs were obtained from FooDB, a database of food constituents (43). First, the pre-trained OptNCMiner using the base dataset was transfer-learned based on chemical-protein interaction data for the target proteins assigned. Second, 65,038 NCs from FooDB were then provided as input to the transfer-learned OptNCMiner to predict protein targets. Third, the transfer-learned OptNCMiner was used for few-shot learning of three target proteins assigned for few-shot learning. In the same manner, NCs from FooDB were provided as input to predict chemical-protein interactions for the three proteins. OptNCMiner achieved a high recall value for all proteins (Additional file 1: Table S4).

To visualize the network of relationships between T2DM complications and NCs of food origin (Fig. 4), a standardized process was followed. First, chemical-protein pairs with a score above 0.5 were selected. Of these pairs, NCs with the highest number of target proteins were selected. Here, 102 NCs modulating 5 different target proteins were chosen. The 102 NCs were matched with their food sources, which added up to 680, as most NCs are found in multiple food sources (see Additional file 3). For ease of visualization, a food source that contains the highest number of selected NCs is shown in Fig. 4. The relationships between T2DM complications, target proteins, NCs, and foods were visualized using connected edges. In identifying food sources to fight T2DM complications, the figure shows that ginger contains 31 NCs that modulates 8 different target proteins associated with T2DM complications.

Ginger is a herbaceous plant that has a long history of use in traditional medicines and foods. The root of the plant contains a vast array of NCs which are responsible for a wide range of biological activities, including anti-diabetic effects, gastrointestinal protection, anti-cancer effects, cardiovascular protection, and the prevention of obesity (44, 45). Among the 160 identified NCs present within ginger, phenolic and terpene compounds, such as gingerols and shogaols, have been widely investigated for their pungent stringency, abundance, and multiple health benefits (44, 46, 47). Derivatives of gingerols and shogaols represented a large proportion of ginger-derived NCs identified by OptNCMiner. Interestingly, OptNCMiner

predicted 6-gingerol as an NC that ameliorates T2DM complications, by modulating 5 different proteins: TLR4, PI3K, YAP, PPARα, and GPR120.

TLR4 is a receptor protein with multiple physiological functions, and has been implicated in the weakening of ocular surfaces and corneal nerves, leading to diabetic keratopathy. TLR4 binds to high-mobility group box 1 protein that activates the NF-κB pathway, leading to inflammation in the cornea (48). Interestingly, it has been reported that gingerols including 6-gingerol and 6-shogaol, inhibit activation of the TLR4 signaling pathway, a finding similar to our predictions (49). Activation of PI3K signaling protects against cardiomyopathy, which is characterized by adverse remodeling of the heart, diastolic dysfunction, fibrosis, and apoptosis (50). YAP, a transcriptional regulator protein, is involved in a key pathway in diabetic cardiomyopathy pathogenesis (51). 6-gingerol is known to protect against hypoxia-induced myocardial injury by activating the PI3K/Akt pathway (52, 53). Furthermore, it has been reported that 6-gingerol treatment inhibits YAP activation by increasing its phosphorylation and preventing translocation into the nucleus (54). Although it remains to be determined whether PPARα is a direct molecular target of 6-gingerol, it is known that 6-gingerol activates the glucagon-like peptide-1 mediated insulin secretion pathway (52), which inhibits PPARα-mediated lipid accumulation and toxicity in cardiomyopathy (55). Additionally, the association between GPR120 and 6-gingerol is also unknown. However, a previous study has reported that *in silico* docking of gingerol to GPR120 yields negative binding energy, suggesting there is a high likelihood of the compound directly binding to GPR120 (56). Activation of GPR120 inhibits TAK1 phosphorylation, which is associated with the induction of proinflammatory responses including TNF-α, IL-6, and COX-2 via NF-κB and IKKβ activation. These proinflammatory and fibrosis signaling cascades lead to the development of diabetic nephropathy. Li and colleagues have reported that 6-gingerol suppresses NF-κB and IKKβ activation as well as the production of NF-κB-dependent inflammatory cytokines in vivo (57). Thus, it is reasonable to hypothesize that GPR120 is a direct molecular target of 6-gingerol for protective effects against diabetic nephropathy. From the results of previous studies mentioned, 6-gingerol is a potent NC that ameliorates T2DM complications. OptNCMiner has correctly predicted the previously known targets of 6-gingerol as well as potential targets that are not yet known, exemplifying its capacity to predict NCs relevant for specific diseases.

An important recommendation for users of OptNCMiner and room for improvement for better performance of OptNCMiner

OptNCMiner was developed to overcome the limitations faced by existing NC identification methods and to support the effective discovery of novel NCs based on the consideration of multi-target interactions. OptNCMiner was built in an SNN structure to enable learning of hidden structural characteristics that determine chemical-protein interactions. The pair-wise input generation and few-shot learning characteristics of OptNCMiner enable this learning even with small datasets. OptNCMiner generates embedding vectors that place compounds with similar target protein interaction closely in a vector space. Thus, OptNCMiner can be used along with other NC discovery methods such as chemical-chemical

interaction prediction (58), toxicity prediction (59), and transcription response prediction (30), to better examine functional NCs from different angles.

However, there remains room for improvement for better performance of OptNCMiner. One factor limiting the performance of the program is the complexity of NC biological activity in the human body. Slight differences in physicochemical properties can affect the absorption and distribution of NCs, which alters the interaction between NCs and target proteins. For further development, the physicochemical properties of NCs as well as the cellular location of target proteins should be considered in model input and model training. Although *in silico* docking has validated predicted false positives as true positives, further validation with *in vitro* and *in vivo* studies will verify the accuracy of OptNCMiner. This effort of further validation would also clarify the synergistic or unexpected side-effects of multi-target modulating NCs discovered by OptNCMiner.

One important recommendation for users of our model is the careful selection of target proteins for NC discovery. In the example of T2DM complications, only proteins ameliorating the three complications were considered target proteins. However, to identify NCs that only regulate the desired target protein, in practice, possible off-target proteins may be affected based on background knowledge. OptNCMiner can also be used in conjunction with other programs to support holistic NC discovery, such as programs that predict the absorption and distribution of NCs after ingestion.

Conclusion

In this study, a novel SNN model called OptNCMiner was built to predict multiple target proteins of NCs. Trained to understand similarities between paired fingerprint vectors, OptNCMiner can predict chemical-protein interactions even for proteins with limited and unbalanced training data. We have demonstrated that OptNCMiner can successfully adapt to training data of various sizes and can predict novel chemical-protein interactions. Furthermore, as a use-case scenario, OptNCMiner was used to predict a natural source and its NCs for the potential treatment of T2DM complications. With a careful selection of target protein candidates, OptNCMiner is a powerful tool to predict novel NCs that modulate specific target proteins to elicit the desired bioactivity.

Abbreviations

AUROC: area under the receiver operation characteristic; BP: boiling point; DNN: deep neural network; GPR120: G protein coupled receptor 120; HL: Henry's Law constant; KOA: octanol-air partition coefficient; LogD: pH-dependent lipid-aqueous partition coefficient; LogP: octanol-water partition coefficient; MLP: multi-layer perceptron; MP: melting point; NC: natural compound; NP: natural product; Nrf-2: nuclear factor erythroid 2-related factor; PCA: principle component analysis; PI3K: phosphoinositide 3-kinase; pKa: acid dissociation constant; PKCβ: protein kinase C β; PPARα: peroxisome proliferator activated receptor α; RT: retention time; SGLT2: sodium-glucose cotransporter 2; SNN: Siamese neural network; TLR4: toll-like receptor 4; VP: vapor pressure; WS: water solubility; YAP: Yes-associated protein

Declarations

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article, additional files, and GitHub at https://github.com/phytoai/OptNCMiner.

Ethics approval and consent to participate

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Funding

This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2019R1I1A1A01060063), and also supported by the K-BIO KIURI Center Program through the Ministry of Science and ICT (2020M3H1A1073304), and the Brain Korea 21 Plus Program of the Department of Agricultural Biotechnology, Seoul National University, Republic of Korea.

Authors' contributions

SHS performed all data analysis and wrote the first draft of the manuscript. SMO contributed to the data analysis and verified the calculations. SHS, HY and SMO developed the original idea and guided the data analysis and presentation of results. SHS, HY, KWL and JHYP contributed to the final manuscript. HY, KWL provided financial support for the study. All authors read and approved the final manuscript.

Acknowledgements

We would like to thank Hee Jae Kim and Kyung Eun Lee for their help on data preparation and generation.

Consent for publication

Not applicable.

References

1. Gonzalez-Manzano S, Duenas M. Applications of Natural Products in Food. Foods. 2021;10(2).
2. Atanasov AG, Zotchev SB, Dirsch VM, Supuran CT, Taskforce INPS. Natural products in drug discovery: advances and opportunities. Nat Rev Drug Discov. 2021;20(3):200–16.

3. Patra JK, Das G, Lee S, Kang SS, Shin HS. Selected commercial plants: A review of extraction and isolation of bioactive compounds and their pharmacological market value. Trends Food Sci Tech. 2018;82:89–109.
4. Sorokina M, Steinbeck C. Review on natural products databases: where to find data in 2020. J Cheminformatics. 2020;12(1).
5. Ahmed J, Preissner S, Dunkel M, Worth CL, Eckert A, Preissner R. SuperSweet-a resource on natural and artificial sweetening agents. Nucleic Acids Res. 2011;39:D377-D82.
6. Sparks TC, Wessels FJ, Lorsbach BA, Nugent BM, Watson GB. The new age of insecticide discovery-the crop impact of natural products. Pestic Biochem Phys. 2019;161:12–22.
7. Dunkel M, Schmidt U, Struck S, Berger L, Gruening B, Hossbach J, et al. SuperScent-a database of flavors and scents. Nucleic Acids Res. 2009;37:D291-D4.
8. Vontzalidou A, Chaita E, Aligiannis N, Makropoulou M, Kalpoutzakis E, Termentzi A, et al. Evaluation of natural products as potential cosmetic agents with tyrosinase inhibition activity. Planta Med. 2012;78(11):1066-.
9. Schuster VTaD. Computational Studies on Natural Products for the Development of Multi-target Drugs. Methods in Pharmacology and Toxicology. 2018:187–201.
10. Loscher W. Single-Target Versus Multi-Target Drugs Versus Combinations of Drugs With Multiple Targets: Preclinical and Clinical Evidence for the Treatment or Prevention of Epilepsy. Front Pharmacol. 2021;12.
11. Cote B, Carlson LJ, Rao DA, Alani AWG. Combinatorial resveratrol and quercetin polymeric micelles mitigate doxorubicin induced cardiotoxicity in vitro and in vivo. J Control Release. 2015;213:128–33.
12. Cheng YT, Yang CC, Shyur LF. Phytomedicine-Modulating oxidative stress and the tumor microenvironment for cancer therapy. Pharmacological Research. 2016;114:128–43.
13. Pearson HE, Iida M, Orbuch RA, McDaniel NK, Nickel KP, Kimple RJ, et al. Overcoming Resistance to Cetuximab with Honokiol, A Small-Molecule Polyphenol. Mol Cancer Ther. 2018;17(1):204–14.
14. Britton RG, Kovoor C, Brown K. Direct molecular targets of resveratrol: identifying key interactions to unlock complex mechanisms. Ann Ny Acad Sci. 2015;1348:124–33.
15. Singh AP, Singh R, Verma SS, Rai V, Kaschula CH, Maiti P, et al. Health benefits of resveratrol: Evidence from clinical studies. Med Res Rev. 2019;39(5):1851–91.
16. Schuster D, Laggner C, Langer T. Why drugs fail - A study on side effects in new chemical entities. Curr Pharm Design. 2005;11(27):3545–59.
17. Cereto-Massague A, Ojeda MJ, Valls C, Mulero M, Garcia-Vallve S, Pujadas G. Molecular fingerprint similarity search in virtual screening. Methods. 2015;71:58–63.
18. Protein Targeting Compounds. Prediction, Selection and Activity of Specific Inhibitors. Anticancer Res. 2016;36(8):4373.
19. Munteanu CR, Fernandez-Blanco E, Seoane JA, Izquierdo-Novo P, Rodriguez-Fernandez JA, Prieto-Gonzalez JM, et al. Drug Discovery and Design for Complex Diseases through QSAR Computational

Methods. Curr Pharm Design. 2010;16(24):2640–55.

20. Sajadi SZ, Chahooki MAZ, Gharaghani S, Abbasi K. AutoDTI plus plus: deep unsupervised learning for DTI prediction by autoencoders. Bmc Bioinformatics. 2021;22(1).
21. Lee I, Keum J, Nam H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. Plos Comput Biol. 2019;15(6).
22. Rifaioglu AS, Atas H, Martin MJ, Cetin-Atalay R, Atalay V, Dogan T. Recent applications of deep learning and machine intelligence on in silico drug discovery: methods, tools and databases. Brief Bioinform. 2019;20(5):1878–912.
23. Fresnais L, Ballester PJ. The impact of compound library size on the performance of scoring functions for structure-based virtual screening. Brief Bioinform. 2021;22(3).
24. Vamathevan J, Clark D, Czodrowski P, Dunham I, Ferran E, Lee G, et al. Applications of machine learning in drug discovery and development. Nat Rev Drug Discov. 2019;18(6):463–77.
25. Jane Bromley IG, Yann LeCun, Eduard Sickinger and Roopak Shah. Signature Verification using a "Siamese" Time Delay Neural Network. NIPS'93: Proceedings of the 6th International Conference on Neural Information Processing Systems. 1994:737–44.
26. Thiolliere R, Dunbar E, Synnaeve G, Versteegh M, Dupoux E. A Hybrid Dynamic Time Warping-Deep Neural Network Architecture for Unsupervised Acoustic Modeling. 16th Annual Conference of the International Speech Communication Association (Interspeech 2015), Vols 1–5. 2015:3179-83.
27. Chopra S, Hadsell R, LeCun Y. Learning a similarity metric discriminatively, with application to face verification. Proc Cvpr Ieee. 2005:539–46.
28. Berlemont S, Lefebvre G, Duffner S, Garcia C. Class-balanced siamese neural networks. Neurocomputing. 2018;273:47–56.
29. Ruffieux S, Lalanne D, Mugellini E. ChAirGest - A Challenge for Multimodal Mid-Air Gesture Recognition for Close HCI. Icmi'13: Proceedings of the 2013 Acm International Conference on Multimodal Interaction. 2013:483-8.
30. Jeon M, Park D, Lee J, Jeon H, Ko M, Kim S, et al. ReSimNet: drug response similarity prediction using Siamese neural networks. Bioinformatics. 2019;35(24):5249–56.
31. Sun JM, Jeliazkova N, Chupakin V, Golib-Dzib JF, Engkvist O, Carlsson L, et al. ExCAPE-DB: an integrated large scale dataset facilitating Big Data analysis in chemogenomics. J Cheminformatics. 2017;9.
32. Tran-Nguyen VK, Jacquemard C, Rognan D. LIT-PCBA: An Unbiased Data Set for Machine Learning and Virtual Screening. J Chem Inf Model. 2020;60(9):4263–73.
33. Steinbeck C, Hoppe C, Kuhn S, Floris M, Guha R, Willighagen EL. Recent developments of the Chemistry Development Kit (CDK) - An open-source Java library for chemo- and bioinformatics. Curr Pharm Design. 2006;12(17):2111–20.
34. Kingma DP, Ba, J. Adam: a method for stochastic optimization. arXiv:14126980 2014.

35. Cai CJ, Wang SW, Xu YJ, Zhang WL, Tang K, Ouyang Q, et al. Transfer Learning for Drug Discovery. J Med Chem. 2020;63(16):8683–94.
36. Chu Y, Shan X, Chen T, Jiang M, Wang Y, Wang Q, et al. DTI-MLCD: predicting drug-target interactions using multi-label learning with community detection method. Brief Bioinform. 2021;22(3).
37. Mansouri K, Grulke CM, Judson RS, Williams AJ. OPERA models for predicting physicochemical properties and environmental fate endpoints. J Cheminform. 2018;10(1):10.
38. Elijah Cole OMA, Titouan Lorieul, Pietro Perona, Dan Morris, Nebojsa Jojic. Multi-Label Learning from Single Positive Labels. Computer Vision and Pattern Recognition. 2021.
39. Shin WH, Kim JK, Kim DS, Seok C. GalaxyDock2: protein-ligand docking using beta-complex and global optimization. J Comput Chem. 2013;34(30):2647–56.
40. Baek M, Shin WH, Chung HW, Seok C. GalaxyDock BP2 score: a hybrid scoring function for accurate protein-ligand docking. J Comput Aided Mol Des. 2017;31(7):653–66.
41. Shahcheraghi SH, Aljabali AAA, Al Zoubi MS, Mishra V, Charbe NB, Haggag YA, et al. Overview of key molecular and pharmacological targets for diabetes and associated diseases. Life Sci. 2021;278.
42. Gilson MK, Liu TQ, Baitaluk M, Nicola G, Hwang L, Chong J. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res. 2016;44(D1):D1045-D53.
43. Centre TMI. FooDB. Canada: The Metabolomics Innovaion Centre; 2017.
44. Zhang M, Zhao R, Wang D, Wang L, Zhang Q, Wei S, et al. Ginger (Zingiber officinale Rosc.) and its bioactive components are potential resources for health beneficial agents. Phytother Res. 2021;35(2):711–42.
45. Anh NH, Kim SJ, Long NP, Min JE, Yoon YC, Lee EG, et al. Ginger on Human Health: A Comprehensive Systematic Review of 109 Randomized Controlled Trials. Nutrients. 2020;12(1).
46. Wohlmuth H, Leach DN, Smith MK, Myers SP. Gingerol content of diploid and tetraploid clones of ginger (Zingiber officinale Roscoe). J Agric Food Chem. 2005;53(14):5772–8.
47. Mao QQ, Xu XY, Cao SY, Gan RY, Corke H, Beta T, et al. Bioactive Compounds and Bioactivities of Ginger (Zingiber officinale Roscoe). Foods. 2019;8(6).
48. Yang W, Feng Y, Zhou J, Cheung OK-W, Cao J, Wang J, et al. A selective HDAC8 inhibitor potentiates antitumor immunity and efficacy of immune checkpoint blockade in hepatocellular carcinoma. Science Translational Medicine. 2021;13(588):eaaz6804.
49. Chen CY, Kao CL, Liu CM. The Cancer Prevention, Anti-Inflammatory and Anti-Oxidation of Bioactive Phytochemicals Targeting the TLR4 Signaling Pathway. Int J Mol Sci. 2018;19(9).
50. Huynh K, Bernardo BC, McMullen JR, Ritchie RH. Diabetic cardiomyopathy: mechanisms and new treatment strategies targeting antioxidant signaling pathways. Pharmacol Ther. 2014;142(3):375–415.
51. Xu S, Zhang H, Liu T, Wang Z, Yang W, Hou T, et al. 6-Gingerol suppresses tumor cell metastasis by increasing YAP(ser127) phosphorylation in renal cell carcinoma. J Biochem Mol Toxicol.

2021;35(1):e22609.
52. Xu T, Qin G, Jiang W, Zhao Y, Xu Y, Lv X. 6-Gingerol Protects Heart by Suppressing Myocardial Ischemia/Reperfusion Induced Inflammation via the PI3K/Akt-Dependent Mechanism in Rats. Evid Based Complement Alternat Med. 2018;2018:6209679.
53. Ren Q, Zhao S, Ren C. 6-Gingerol protects cardiocytes H9c2 against hypoxia-induced injury by suppressing BNIP3 expression. Artif Cells Nanomed Biotechnol. 2019;47(1):2016–23.
54. Samad MB, Mohsin M, Razu BA, Hossain MT, Mahzabeen S, Unnoor N, et al. [6]-Gingerol, from Zingiber officinale, potentiates GLP-1 mediated glucose-stimulated insulin secretion pathway in pancreatic beta-cells and increases RAB8/RAB10-regulated membrane presentation of GLUT4 transporters in skeletal muscle to improve hyperglycemia in Lepr(db/db) type 2 diabetic mice. BMC Complement Altern Med. 2017;17(1):395.
55. Wu L, Wang K, Wang W, Wen Z, Wang P, Liu L, et al. Glucagon-like peptide-1 ameliorates cardiac lipotoxicity in diabetic cardiomyopathy via the PPARalpha pathway. Aging Cell. 2018;17(4):e12763.
56. Chinthakunta N, Cheemanapalli S, Chinthakunta S, Anuradha CM, Chitta SK. A new insight into identification of in silico analysis of natural compounds targeting GPR120. Network Modeling Analysis in Health Informatics and Bioinformatics. 2018;7(1):8.
57. Li Y, Xu B, Xu M, Chen D, Xiong Y, Lian M, et al. 6-Gingerol protects intestinal barrier from ischemia/reperfusion-induced damage via inhibition of p38 MAPK to NF-kappaB signalling. Pharmacol Res. 2017;119:137–48.
58. Kwon S, Yoon S. End-to-End Representation Learning for Chemical-Chemical Interaction Prediction. Ieee Acm T Comput Bi. 2019;16(5):1436–47.
59. Mayr A, Klambauer G, Unterthiner T, Hochreiter S. DeepTox: Toxicity Prediction using Deep Learning. Front Env Sci-Switz. 2016;3.

Figures

Figure 1

OptNCMiner model flowchart

Figure 2

The distribution of (a) physicochemical properties; and (b) chemical structures in the base dataset, transfer learning dataset, and limited-shot learning dataset.

Figure 3

In silico docking score for false positives from the (a) few-shot learning dataset; and (b) two molecular docking results for compound-protein interactions with lowest *in silico* docking score (highest binding affinity).

Figure 4

OptNCMiner predicts that ginger contains 33 NCs that regulate 8 different target proteins associated with T2DM complications.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.docx
- Additionalfile2.csv
- Additionalfile3.csv