Journal of Cheminformatics

DATABASE

Open Access

TCMSID: a simplified integrated database for drug discovery from traditional chinese medicine

Liu-Xia Zhang[1†], Jie Dong[2†], Hui Wei[2], Shao-Hua Shi[2,3], Ai-Ping Lu[3*], Gui-Ming Deng[1*] and Dong-Sheng Cao[2,3*]

Abstract

Traditional Chinese Medicine (TCM) has been widely used in the treatment of various diseases for millennia. In the modernization process of TCM, TCM ingredient databases are playing more and more important roles. However, most of the existing TCM ingredient databases do not provide simplification function for extracting key ingredients in each herb or formula, which hinders the research on the mechanism of actions of the ingredients in TCM databases. The lack of quality control and standardization of the data in most of these existing databases is also a prominent disadvantage. Therefore, we developed a Traditional Chinese Medicine Simplified Integrated Database (TCMSID) with high storage, high quality and standardization. The database includes 499 herbs registered in the Chinese pharmacopeia with 20,015 ingredients, 3270 targets as well as corresponding detailed information. TCMSID is not only a database of herbal ingredients, but also a TCM simplification platform. Key ingredients from TCM herbs are available to be screened out and regarded as representatives to explore the mechanism of TCM herbs by implementing multi-tool target prediction and multilevel network construction. TCMSID provides abundant data sources and analysis platforms for TCM simplification and drug discovery, which is expected to promote modernization and internationalization of TCM and enhance its international status in the future. TCMSID is freely available at https://tcm.scbdd.com.

Keywords: Ingredient database, Chinese medicine, Key ingredients, Multi-tool target prediction

Introduction

Traditional Chinese Medicine (TCM) has played a vital role in extensively treating various diseases for thousands of years in China. In virtue of its exact curative effect, TCM is still used to maintain human health until today and has received worldwide attention. Virtually, TCM restores the human body to normal physiological condition by perturbing the human dysfunctional network through its ingredients, which is consistent with the role of western medicine, despite having a fundamental theory that is very different from that of contemporary western medicine. For example, ephedrine and pseudoephedrine, etc., are regarded as the major ingredients of Ephedra Decoction to treat colds and relieve cough and asthma. Moreover, given the characteristics of multi-ingredient and multi-target, TCM is actually more in line with the trend of current combination therapy of multi-drugs which holds great potential for treating various intractable diseases, such as malignant tumors and cardiovascular diseases rather than the prior concept of 'one drug-one gene-one disease' [1].

Herbal ingredients, as the footstone of the TCM, have long been recognized as an ideal starting point for

†Liu-Xia Zhang and Jie Dong–Joint first authors

*Correspondence: aipinglu@hkbu.edu.hk; guimingd1004@163.com; oriental-cds@163.com

[1] The First Hospital of Hunan University of Chinese Medicine, Changsha 410007, Hunan, People's Republic of China
[3] Advancing Translational Medicine in Bone and Joint Diseases, School of Chinese Medicine, Hong Kong Baptist University, Hong Kong SAR, People's Republic of China
Full list of author information is available at the end of the article

Table 1 Present-existing major databases of TCM ingredients/natural products and their basic characteristics

Database name	TCM	Ingredients	Target	ADME/T	Experimental bioactivity data	Quality control	Website
TCM@Taiwan	453	> 20,000	N/A	N/A	N/A	N/A	http://tcm.cmu.edu.tw
HIT	1300	586	1301	N/A	N/A	N/A	http://lifecenter.sgst.cn/hit/
TCMSP	499	29,384	3311	Yes	N/A	N/A	http://tcmspw.com/news.php
TCMID2.0	8159	43,413	17,521	N/A	N/A	Less	http://www.megabionet.org/tcmid/
ETCM	403	7274	7603	Yes	N/A	Yes	http://www.nrc.ac.cn:9090/ETCM/
NPASS	N/A	35,032	5863	N/A	Yes	N/A	http://bidd2.nus.edu.sg/NPASS/
NPACT	N/A	1574	284	Yes	Yes	N/A	http://crdd.osdd.net/raghava/npact/

molecular design owing to their broad chemical structural diversity and high selectivity [2]. It is estimated that no less than 50% approved small-molecular clinical drugs in the world are directly or indirectly derived from herbal ingredients [3, 4]. Artemisinin extracted from *Artemisia annua* is a well-known example to treat malaria. However, it is the multi-ingredient and multi-target characteristics of the TCM that make the mechanism of action remain exclusive, thus hindering its application and modernization [5, 6].

The common strategy for TCM mechanism exploration is to map all the ingredients of a prescription to the ingredient-target network and infer probable mechanisms of the prescription from enriched pathways [7–9]. It is well known that there are tens of thousands of ingredients in TCM, even the whole chemical information of an herb/prescription from TCM itself can be treated as a database of natural products; however, most of the ingredients are ineffective and redundant and rarely or even do not exert effective pharmacological responses owing to their low content and activity. Moreover, the reliability of the target is critical to revealing the mechanism. The rough strategy used above only depicts an impractical blueprint instead of deciphering the real mechanisms of TCM, which makes the inherently complex mechanism more confusing and hinders the study of mechanism. Therefore, the simplification step for TCM ingredients is imperative for clarifying mechanisms and promoting the development of TCM.

Over the past decade, with the emergence of several herbal ingredient databases, such as TCM@Taiwan [10], TCMSP [11], TCMID [12], etc., acquisition of TCM ingredient information has no longer been limited to articles, which have largely facilitated the subsequent mechanism study and TCM modernization process. However, these databases inevitably have several certain shortcomings. Despite the fact that TCM@Taiwan increased the number of his component records to 6100 in 2014, it still lacks TCM classification and compound-target data. HIT is a target-focused database with limited ingredients data. TCMSP is a comprehensive database containing a high number of records for TCM ingredients, targets, diseases, and even ADME, however, it does not contain TCM classification and quality control methods. The number of records for each item of TCMID is even greater than that of TCMSP and TCMSID. Even though it collected MS data of 3895 TCM ingredients to conduct quality control, TCMID 2.0 did no herb classification. ETCM is another TCM database with comprehensive data types, which included herb classification, ingredient data, target data, disease data, quality control, network display and even TCM pictures. However, the number of entries of nearly every data type is less than TCMSP, TCMID 2.0 and TCMSID respectively. NPASS and NPACT are characterized by the experimental data load they included, yet the two databases included no TCM classification, ADME/T information and quality control method. While each of these databases has its own advantages and complements each other, TCMSID includes almost all of the advantages. More importantly, TCMSID provided simplification function for extracting key ingredients in each herb or formula, which is the unique and significant trait of the database [13]. Although some databases provide mechanism analysis function, most of them reveal the pharmacological mechanism of the whole ingredients of an herb/prescription based on a rough mapping strategy, which only results in ambiguous mechanism. Present-existing major databases of TCM ingredients/natural products and their basic characteristics are shown in Table 1.

In this study, we developed TCMSID, a Traditional Chinese Medicine Simplified Integrated Database (https://tcm.scbdd.com). TCMSID is a database of high storage, high quality and standard, which is specifically manifested in the following aspects: (1) integration of 499 TCM herbs and 20,015 unique herbal ingredients, which largely compensates for the prior-existing databases; (2) the adjunction with multiple aspects of comprehensive information for each ingredient contained in TCM herbs, including significance degree, ADME/T-related

Table 2 Data source and volume of TCMSID

Item	Data source	Amount of data
TCM herbs	TCMSP, SymMap	499
Total ingredients	Literature mining, TCMSP, SymMap	20,015
Herb-ingredient associations	Literature mining, TCMSP, SymMap	50,053
Focused targets	SEA, Swiss targetprediction, HitPickV2, PPB, PPB2, ChEMBL	2390
Drugs	DrugBank	10,487

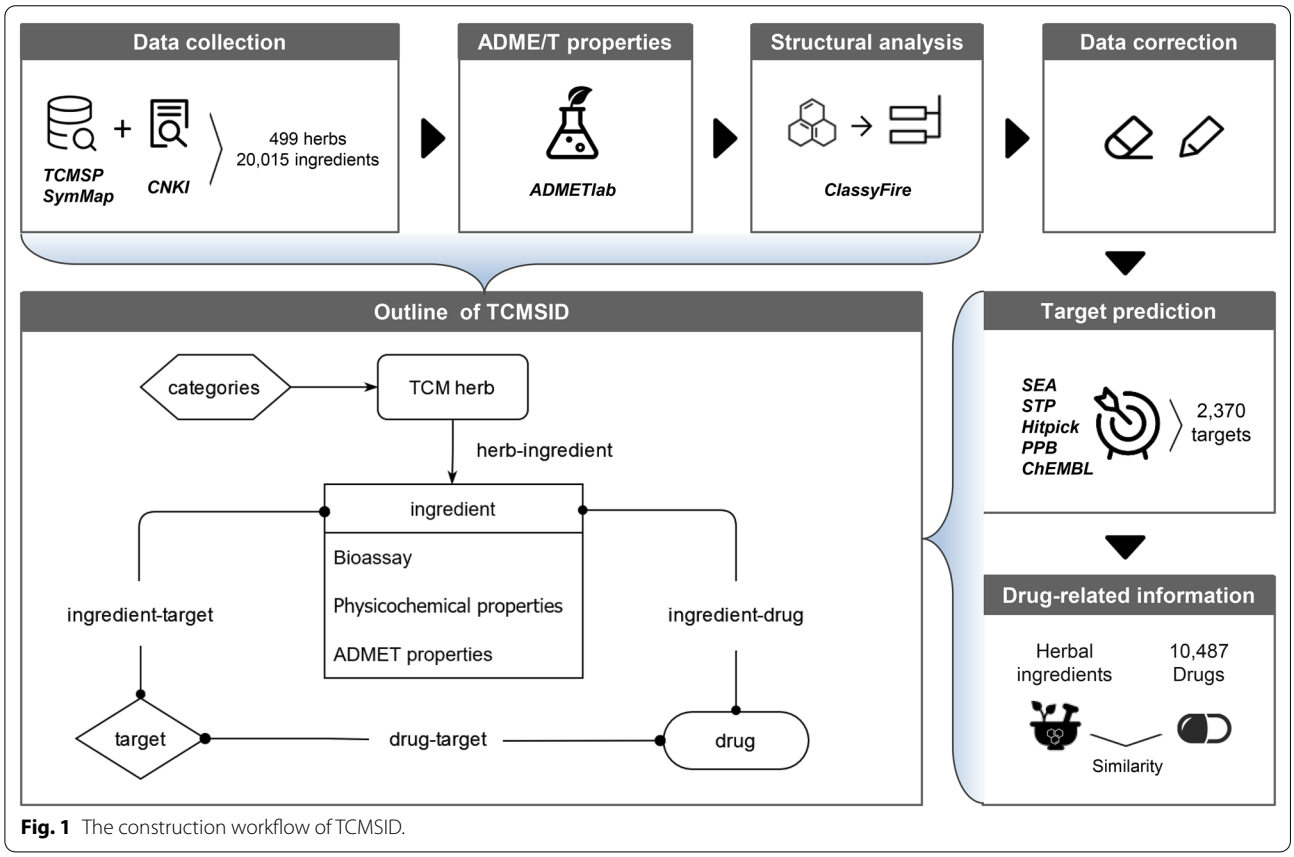

Fig. 1 The construction workflow of TCMSID.

properties, structural classification and reliability; (3) the incorporation of reliable potential targets predicted by multiple target-prediction platforms for each ingredient; (4) the provision of bioassay data for herbal ingredients, which can be used to study hidden activity-related information using cheminformatics methods; (5) the establishment of an herb-component-target-drug multi-level interaction network of TCM for deeper study of the mechanism of actions. A summary of the construction workflow of TCMSID is shown in Fig. 1.

Most importantly, TCMSID is not only a repository of TCM ingredients available for query purpose, but also an analysis platform to facilitate clarifying the mechanism of actions. Key ingredients can be screened out as representative ingredients for pharmacological activities exerted by respective TCM herbs and used for multi-tool target prediction to obtain reliable targets to finally clarify the whole mechanism. Also, it provides data analysis and visualization of TCM related information on the network level. The data volume of TCMSID is summarized in Table 2.

Implementation and functionalities

TCMSID is composed of five fields, including TCM categories, TCM herb, ingredient, target and drug (Fig. 1). Detailed information of each field was integrated from other relevant databases, text mining of published articles and prediction tools such as ADMETlab [14, 15].

In virtue of these interrelated fields, users can conduct a query relying on keywords of any field as an entry point and retrieve relevant information as needed based on the corresponding links. To conduct TCM simplification and mechanism analysis, representative key ingredients of an herb, which exert the pharmacological action of the herb, are available to be screened out. The identification method is based on the detailed information about the ingredients, mainly including significance degree, ADME/T, physicochemical properties, structural reliability, and structural characteristics. Meanwhile, a multilevel functional network can be built through the resulting key ingredients, the reliable targets of the key ingredients and the similar-drug-related information of the key ingredients. This network bridges the gap between TCM and modern medicine. Next, we will elaborate on the detailed information and acquisition process for constructing TCMSID concluded in Fig. 1.

Data processing and implementation

Herbal ingredients

To ensure the high storage of the database, 499 frequently used and approved TCM herbs were collected from the Pharmacopoeia of the People's Republic of China (2015 version). It is well known that a TCM herb is more likely contains hundreds of compounds and can even be regarded as a small compound library, however, not all the ingredients contained are pharmacologically active. Herein, to extract the major active ingredients of TCM herbs, more than 1500 Chinese articles researching these TCM herbs were retrieved from China National Knowledge Infrastructure (CNKI) (http://www.cnki.net/), since TCM was widely used and researched in China and the related research results were mainly published in Chinese as well. The ingredients with high content and activity were extracted through manual mining of these literatures. The herbal ingredients from those publications, as well as from other related web-based databases including TCMSP and SymMap [16] form the data foundation of TCMSID. The significance degree ranges from 0 to 2, the smaller the number, the higher the significance degree. The three numbers are assigned by the bioactivity data and the minimum volume of a compound per unit to exert pharmacological effect according to the referred literature and the Pharmacopoeia of the People's Republic of China (2015 version), respectively. For the compound that satisfies the criteria of both bioactivity and minimum volume per unit, we assign the degree of significance the value of 0; for the compound that satisfies the criteria of either bioactivity or minimum volume per unit, we assign the degree of significance the value of 1; and for the compound that fails to satisfy the criteria of both bioactivity and minimum volume per unit, we assign the degree of significance the value of 2. The data of bioactivity and minimum volume per unit for the 499 TCM herbs was manually collected from the Pharmacopoeia and literature, following which the significance degree values of all the ingredients were assigned according to the aforementioned criteria. Details about those ingredients, such as name, structure etc., were comprehensively retrieved from PubChem (PUG-REST interface) automatically [17, 18], where the structure files in multiple formats (sdf, mol, SMILES etc.) were eventually converted into canonical SMILES using OpenBabel (version 2.4.1). The duplicates were removed according to InChIKey.

ADME/T-related properties

To improve the quality of the database, we conducted an in-depth analysis for each ingredient. First of all, a battery of pivotal drug-likeness properties were computed through our prior work ADMETlab (http://admet.scbdd.com) and ADMETlab 2.0 (https://admetmesh.scbdd.com/), including ADME/T parameters: Caco-2 permeability (Caco-2), Bioavailability (F-30), Plasma Protein Binding (PPB), Blood-Brain Barrier (BBB) Penetration, Half Life ($T_{1/2}$), Clearance (CL), hERG Inhibition (hERG), Human Hepatotoxicity (HHT), drug-likeness (DL), etc. and basic physicochemical parameters: molecular weight (MW), LogP, LogS, etc. Different from most of the property computational tools, ADMETlab and its updated version is an ADME/T evaluation platform, which integrates comprehensive ADME/T properties and basic physicochemical endpoints as many as possible to provide an overall understanding of query compounds and facilitate the drug discovery process.

Before compounds are further investigated in vitro, ADME/T-related properties and basic physicochemical properties are commonly used to provide a fast preliminary filtering. ADME/T-related properties determine whether a molecule will reach the acting site in the body, and how long it will stay in the bloodstream, while basic physicochemical properties closely related to drug-likeness. Property evaluation is nowadays routinely carried out at the early stage of drug discovery to reduce the attrition rate [19, 20], among which the evaluation of pharmacokinetic and physicochemical properties are important prerequisites for filtering key ingredients. As a result, only the major active ingredients that exhibit favorable pharmacokinetic and physicochemical properties can exert potential biological effects.

Ingredient structural classification

To improve the quality of the database, structures of all ingredients were further dissected since the structural characteristic of immense structural diversity is the source of a wide variety of biological activities and the

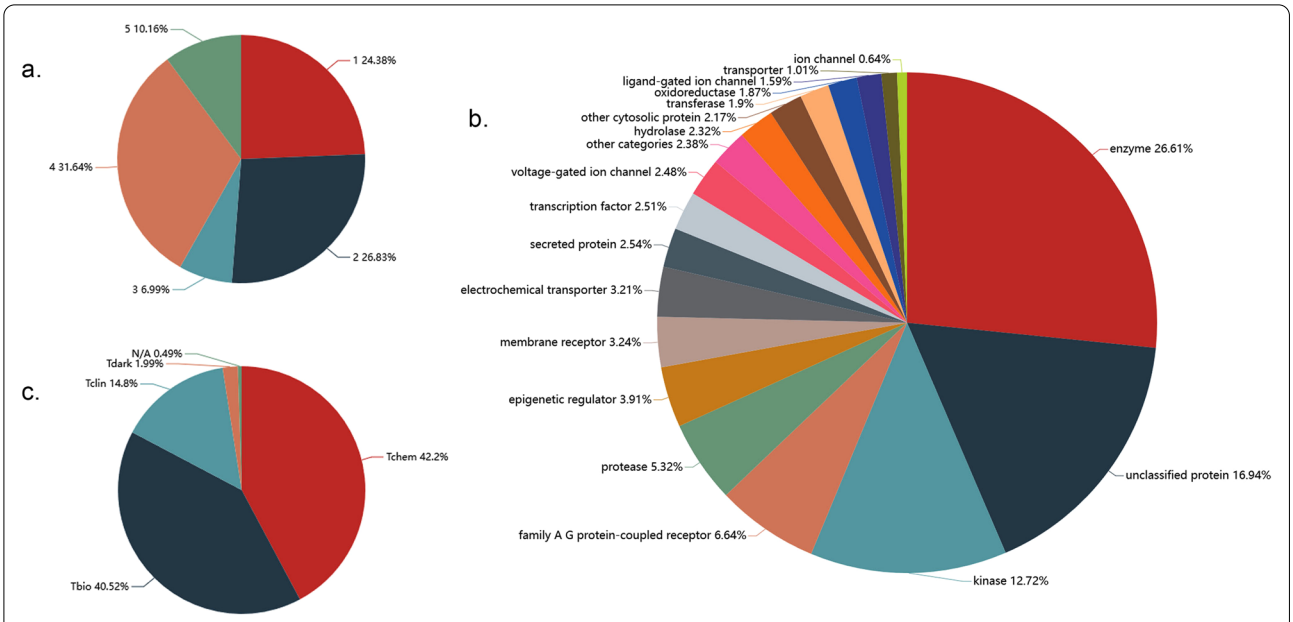

Fig. 2 Quality information of TCMSID ingredients shown in (**a**) the structural reliability evaluation of TCM ingredients. And predicted target information shown in (**b**) the classification of ingredient targets by different homologous families and (**c**) the classification of ingredient targets by development levels

fundamental basis of herbal ingredients for drug design. Herein, ClassyFire web server [21], an automated chemical classification web tool, was used to the refine structural classification of all ingredients layer-by-layer. For instance, matrine, an alkaloid found in plants and a key active ingredient in the herb *Sophora flavescent*, was grouped under the headings of alkaloids and derivatives, lupin alkaloids, and matrine alkaloids.

Ingredient structural reliability evaluation

From the perspective of structural quality, the structural reliability of ingredients can be trustworthy insufficiently due to the diverse data sources, which will fundamentally hinder the TCM research process in a great measure. To evaluate the structural reliability of each ingredient for accurate analysis, the reliability annotations, which indicate the structural quality, were gained by performing structural reliability evaluation using a semi-automated quality checking workflow while keeping the ingredients failed to meet criteria with structural reliability marking [22]. The operation principle of the workflow is to input the chemical name and CAS number of any Chinese medicine ingredient, and then retrieve data from several different ingredient databases such as PubChem and evaluate the quality of the ingredient data by comparing the consistency of the search results obtained by the two searching methods. Here, the structural reliability ranges from 1 to 5, in which 1 to 3 means relatively higher structural reliability with 1 the highest reliability, while 4 stands for unknown reliability, and only 5 means poor reliability. For the chemicals with unknown reliability, we performed additional manual inspection and information correction, and then rescored the corrected chemicals following the workflow. (Fig. 2a).

Ingredient target information

To acquire reliable targets for mechanism exploration, target prediction was performed by implementing and assembling different target prediction tools including SEA [23], SwissTargetPrediction [24], HitpickV2 [25], PPB [26], PPB2 [27] and ChEMBL [28]. We introduced occurrence frequency parameter, which refers to the frequency of targets predicted by different tools. For a given target, the higher the occurrence frequency represents the higher-ranking level. Herein, for each prediction tool, only the top 15 predicted targets were retained according to the occurrence frequency parameter.

Comprehensive information for 3270 target proteins was collected from ChEMBL [29]. Detailed annotation information of all targets is obtained by ID conversion through UniProt [30], which included identification names, functionality description, cross-ref IDs, etc. In addition, target proteins involved in this database were classified into different homologous families through ChEMBL target annotation, such as enzyme and ion channel (Fig. 2b). In the meantime, from a clinical, chemical and biological standpoint, the development level of these targets was divided into Tclin (clinic),

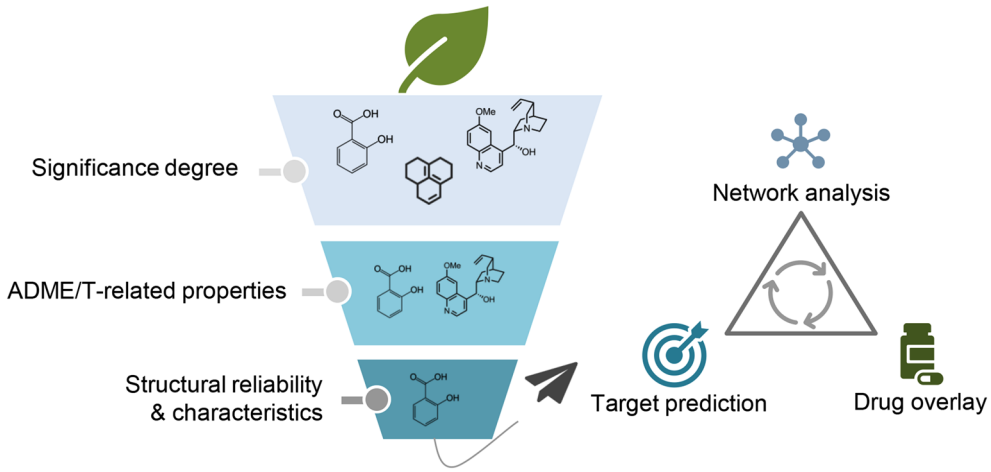

Fig. 3 TCM simplification process of TCMSID

Tchem (chemistry), Tbio (biology) and Tdark (dark genome) using TDL classification scheme developed by Oprea et al. (Fig. 2c) [31].

Drug-related information

To further clarify the knowledge of TCM functions from the modern medicinal point of view, we built the relationship between TCM ingredients and drugs through chemical similarity. The drugs in TCMSID were collected from the DrugBank database [32], which included a total of 10,450 known drugs (containing 3883 FDA-approved drugs), as well as drug-related information including drug names, structures, and drug targets, etc.

Herein, both FCFP6 and ECFP4 fingerprints were adopted to represent all ingredients and drugs since it was previously reported that the circular fingerprint, especially the FCFP6 and ECFP4, show better performance in TCM ingredient similarity search [33, 34]. As a common measure method for 2D similarity, Tanimoto coefficient (*Tc*) was applied to define chemical structural similarity between comparative individuals. Moreover, $Tc = 0.85$ and $Tc = 0.5$ were taken respectively as the thresholds to indicate high and medium similarities between query molecules and drugs. Finally, the structural similarities between comparatives were determined by the intersection of similarity results by comparing the two results and adopting the lower level of classification as the final similarity outcome for the two conflicting results. Calculation of fingerprints and chemical similarity was performed using CDK Fingerprints and Similarity Search node of Knime (version 3.7.2), respectively [35].

Functionalities - mechanism exploration of TCM herbs

To achieve Mechanism exploration of TCM herbs, TCMSID provided TCM simplification for clarifying mechanisms, including two key steps of key ingredients filtering and target identification (Fig. 3). The key ingredients, as the fundamental material basis of TCM, refer to several ingredients that are available to replace a TCM to exert effective pharmacological activity to a certain extent. Key ingredients should have the characteristics of high activity and content. In addition, favorable pharmacokinetic and physicochemical properties should be exhibited to exert potential biological effects. Moreover, given the significant role of molecular structure in pharmacological activity, the structural characteristics and reliability of herbal ingredients should be considered as well. TCMSID provided integrative information for each herbal ingredient, including significance degree, ADME/T and physicochemical properties, structural reliability, structural characteristics, etc. The key ingredients can be filtered in a custom way by setting the threshold range of the above information, according to details of parameters and filtering criteria provided by TCMSID.

Reliable target proteins are the core of mechanism research to promote the modernization of TCM herbs. In recent years, *in-silico* target prediction methods have been regarded as an effective alternative to experimental target identification methods due to its convenience and less time-consuming properties. However, a single target prediction method is more likely leading to inaccurate offset results. It is more beneficial to combine these target prediction methods to take different theoretical foundations into account.

To explore the mechanisms of TCM herbs, the reliable targets of key ingredients can be obtained and aggregated

by carrying out multi-tool target prediction. According to the occurrence frequency parameter and detailed target information provided by TCMSID, the potential targets of TCM herbs to exert pharmacological effects can be screened out as well. In addition, TCMSID provides ingredient-related drug information, such as the therapeutic effects and known targets of the drugs being connected, to bridge the gap between TCM herbs and modern drugs through chemical similarity calculation. Finally, the mechanism of action of herbal ingredients can be inferred according to the multilevel herb-ingredient-target-drug network constructed on the network visualization interface.

Utility and discussion

User interface

To facilitate the use of TCMSID and make it more convenient and faster, we developed a user-friendly interface for the database. The whole database is deployed on the instance of Elastic Compute Service (ECS) of Alibaba cloud. Considering that the database needs to meet multi-user data access and long transaction, the relational database MySQL from Alibaba cloud was used as the database backend. Also, considering that the database operation needs to be integrated with the cheminformatics computing environment, we use Python as the main coding language for architecture development, since Python development environment provides mature data processing and modeling ecosystems. On this basis, we used the most popular Python-based web framework, Django, combined with HTML5, CSS and JavaScript language to develop the front-end visual interface.

The whole user interface of TCMSID consists of modules of 'Home', 'Browse', 'Search', 'Help' and 'Contact'. In the 'browse' module, we show the specific categories of herbs and the structural classification of compounds. The 'browse' module is also the main entrance to the database. By expanding the categories of herbs, we can find specific TCM herb and its detailed information. By clicking on the name of the herb, users can get to the main interface of herbal ingredients. On that page, the specific information of herb and the table of ingredients will be displayed. Ingredient tables support some modern operations, such as retrieval, filtering and comparison, which can be realized just by clicking mouse.

In the table, researchers can carry out the first level of TCM simplification. The active ingredients can be simplified and filtered by setting the basic physicochemical properties, structural reliability and ADME/T properties. By clicking the selected ingredient, the detailed information including the identification, basic physicochemical properties, ADME/T results and target prediction results will be displayed on a new page. On this page, we also provide specific information and statistics of targets. Clicking on each target can lead to more detailed information. Here, researchers can simplify the TCM prescription at the second level. By adding a high reliable predicted target and the ingredient itself into the basket which is always affixed at the right bottom of the page, the platform will automatically calculate and generate the network analysis diagram of herb-component-target-drug relationships. Of course, in the 'browse' module, the users can directly expand and view the subcategories and ingredients according to the structure category. Users can also get information about ingredients by clicking on the names.

In the 'search' module, researchers can carry out general retrieval of ingredients and TCMs according to keywords, and the input supports various types of keywords. In the search result list, users can click on items to display their information. In the 'Help' module, we have organized some tutorials on how to use this platform. These tutorials vividly show how to use the above functionalities and customized different analysis pipelines in the form of video. In addition, in the videos, we also showed how to download data and save the results. However, the full database is not publicly downloadable due to server load considerations and website functional design.

Case study

The TCM formula Erzhi Pill (EZP), composed of *Fructus Ligustri Lucidi* and *Herba Ecliptae*, is one of the frequently used classic prescriptions in China with various pharmaceutical functions such as liver protection and anti-tumor effect [36–39]. However, the mechanism of pharmacological activity exerted by EZP is still unclear, which impeded its application and development. Therefore, implementing TCM simplification to determine the key ingredients and potential targets of EZP, thus further reveal its material basis and mechanism of action are currently urgent problems. Herein, TCM simplification of formula EZP was conducted using TCMSID to illustrate the usage and the validity of the database.

To acquire the key ingredients, a total number of 414 ingredients of EZP were firstly retrieved from TCMSID. With the filtering criteria of Significance degree < 2, Structural reliability < 4, Druglikeness = 1 and Caco-2 > − 5.5, six key ingredients were screened out as the representative of EZP, including oleanolic acid, ursolic acid, salidroside, nuezhenide, specnuezhenide and wedelolactone. After literature research, we found that these six ingredients are key active ingredients of TCM herb *Fructus Ligustri Lucidi* and *Herba Ecliptae* [37–40]. Therefore, we have enough reason to believe that the key ingredients identification process of EZP is accurate and reliable.

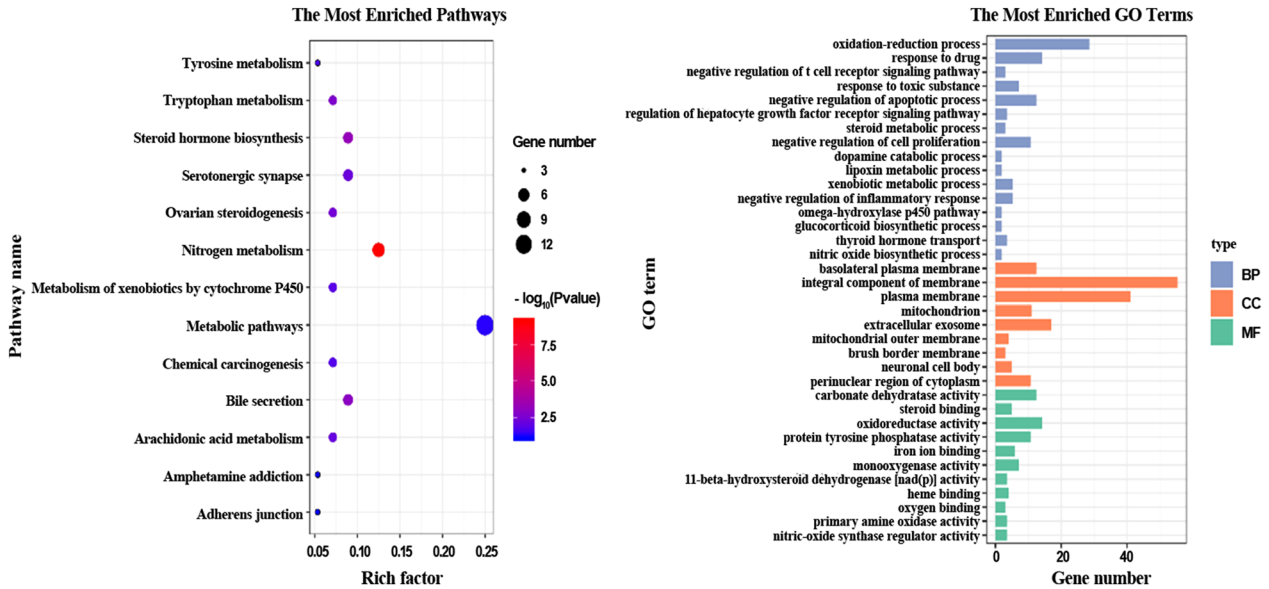

Fig. 4 The top enriched GO terms and KEGG pathways for the potential target proteins of formula EZP through functional enrichment analysis by DAVID 6.8. (BP, MF, and CC represent Biological Process, Molecular Function, and Cellular Component groups of GO, respectively)

To obtain the reliable targets of EZP for mechanism analysis, occurrence frequency = 3 was taken as the identification threshold, which means that only target proteins predicted by at least 3 target prediction tools can be identified. Finally, 56 target proteins that met the condition were picked out as the potential targets. To verify the effectiveness of these target proteins, the enrichment analysis was further implemented by Diversity Visualization Integrated Database (DAVID, version 6.8) [41]. As is shown in Fig. 4, we found that these target proteins were closely related to the liver protection and anti-tumor pharmacological activities of EZP through the top enriched KEGG pathways and GO terms[42], such as hsa00910 Nitrogen metabolism, hsa00140 Steroid hormone biosynthesis, hsa00380 Tryptophan metabolism, GO: 0050860 negative regulation of T cell receptor signaling pathway and GO: 0008285 negative regulation of cell proliferation. According to the target-related information provided by TCMSID, 25 target proteins were further identified as the key targets of pharmacological activities exerted by EZP by reference to target classification and reduction of targets and checking irrelevant function. These key targets were closely related to the genes responsible for occurrence and development of liver diseases and tumors, such as CBR1, PTPN1, NQO1, ALOX12, HSD11B2 and HSP90AA1 (Fig. 5, generated by Cytoscape 3.5.1) [43–49]. In a word, we believed that the TCM simplification and mechanism analysis of EZP can be conveniently achieved through TCMSID.

From the selection of herbs in the formulation, to the screening of key ingredients, to the identification of target prediction targets, users can customize the process in each of these steps according to the parameters they choose. As a result, the constructed EZP—key-ingredient—target visualization network is concise but reliable, which can further clarify the mechanism of action of TCM herbs or formulas. For the prediction results from each module, TCMSID provides download

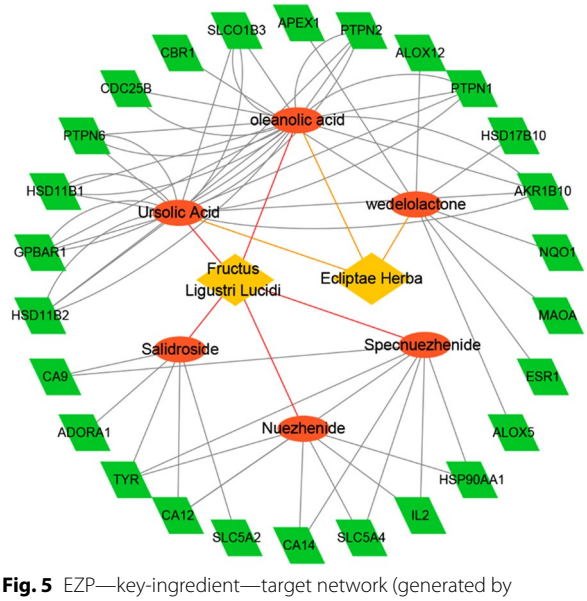

Fig. 5 EZP—key-ingredient—target network (generated by Cytoscape 3.5.1)

options to these analysis results in multiple file formats as well as external links for more detailed information.

Despite of the convenient function of filtering key ingredients of TCM herb or formulas and predicting target proteins potentially connect to these ingredients, the platform still has some limitations require further discussion and improvement. The role of "Jun-Chen-Zuo-Shi" (also known as "sovereign-minister-assistant-courier") classifies prescription TCM herbs by the importance or the role each of them are playing in. The TCM formula data, such as formula component and the Jun-Chen-Zuo-Shi labeling, is of great importance when users search key ingredients herb by herb. By knowing the formula content, especially the herb importance labeling, users can choose to query herbs based their own decision referring to the role of each herb. Data quality is essential for building prediction models. Better computer-aided quality control methods should be further developed for building a database with higher quality data and more accurate prediction functions.

Conclusion

To accelerate the progress of TCM's modernization and standardization, we have presented a Traditional Chinese Medicine Simplified Integrated Database (TCMSID). It is a high storage, high-quality and standardized database, with comprehensive information of ingredients, which largely compensates for the shortcomings of the existing databases. Most importantly, it is not only a data repository just available for information queries, but also a unique mechanism analysis platform for TCM simplification. In short, TCMSID provides data sources and novel research mentality for TCM mechanism research and innovative drug discovery, and it will continue to be developed and updated in the future to promote the modernization and internationalization of TCM.

Abbreviations

TCM: Traditional Chinese Medicine; CNKI: China National Knowledge Infrastructure; PPB: Plasma protein binding; BBB: Blood-brain barrier; CL: Clearance; HHT: Human hepatotoxicity; DL: drug-likeness; MW: molecular weight; ECS: Elastic compute service; EZP: Erzhi Pill.

Acknowledgements

We acknowledge Haikun Xu, and the High-Performance Computing Center of Central South University for support. The study was approved by the university's review board.

Author contributions

APL, GMD and DSC initiated and designed the project; LXZ, JD and HW drafted the manuscript; SHS helped prepare the figures and improve the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by The National Key Research and Development Program of China (2021YFF1201400), National Natural Science Foundation of China (22173118, 22003078), Hunan Provincial Science Fund for Distinguished Young Scholars (2021JJ10068), the Science and Technology Innovation Program of Hunan Province (2021RC4011), Changsha Municipal Natural Science Foundation (kq2014144), Changsha Science and Technology Bureau project (kq2001034), and HKBU Strategic Development Fund project (SDF190402P02).

Availability of data and materials

TCMSID is freely available at https://tcm.scbdd.com.

Declarations

Competing interests

Not applicable.

Author details

[1]The First Hospital of Hunan University of Chinese Medicine, Changsha 410007, Hunan, People's Republic of China. [2]Xiangya School of Pharmaceutical Sciences, Central South University, Changsha 410013, Hunan, People's Republic of China. [3]Advancing Translational Medicine in Bone and Joint Diseases, School of Chinese Medicine, Hong Kong Baptist University, Hong Kong SAR, People's Republic of China.

Received: 19 June 2022 Accepted: 14 December 2022
Published online: 31 December 2022

References

1. Zhou X, Seto SW, Chang D, Kiat H, Razmovski-Naumovski V, Chan K et al (2016) Synergistic effects of Chinese herbal medicine: a comprehensive review of methodology and current research. Front Pharmacol 7:201. https://doi.org/10.3389/fphar.2016.00201
2. Koehn FE, Carter GT (2005) The evolving role of natural products in drug discovery. Nat Rev Drug Discov 4(3):206–220. doi: https://doi.org/10.1038/nrd1657
3. Newman DJ, Cragg GM (2016) Natural Products as sources of new drugs from 1981 to 2014. J Nat Prod 79(3):629–661. doi: https://doi.org/10.1021/acs.jnatprod.5b01055
4. Patridge E, Gareiss P, Kinch MS, Hoyer D (2016) An analysis of FDA-approved drugs: natural products and their derivatives. Drug Discov Today 21(2):204–207. doi: https://doi.org/10.1016/j.drudis.2015.01.009
5. Reker D, Perna AM, Rodrigues T, Schneider P, Reutlinger M, Monch B et al (2014) Revealing the macromolecular targets of complex natural products. Nat Chem 6(12):1072–1078. doi: https://doi.org/10.1038/nchem.2095
6. Rodrigues T, Reker D, Kunze J, Schneider P, Schneider G (2015) Revealing the macromolecular targets of fragment-like natural products. Angew Chem Int Ed Engl 54(36):10516–10520. https://doi.org/10.1002/anie.201504241
7. Li H, Zhao L, Zhang B, Jiang Y, Wang X, Guo Y et al (2014) A network pharmacology approach to determine active compounds and action mechanisms of ge-gen-qin-lian decoction for treatment of type 2 diabetes. Evid Based Complement Alternat Med 2014:495840. doi: https://doi.org/10.1155/2014/495840
8. Xue J, Shi Y, Li C, Song H (2019) Network pharmacology-based prediction of the active ingredients, potential targets, and signaling pathways in compound Lian-Ge granules for treatment of diabetes. J Cell Biochem 120(4):6431–6440. doi: https://doi.org/10.1002/jcb.27933
9. Fan S, Shi X, Wang A, Hou T, Li K, Diao Y (2021) Evaluation of the key active ingredients of 'radix astragali and rehmanniae radix mixture' and related signaling pathways involved in ameliorating diabetic foot ulcers from the perspective of TCM-related theories. J Biomed Inform 123:103904. https://doi.org/10.1016/j.jbi.2021.103904
10. Chen CY (2011) TCM Database@Taiwan: the world's largest traditional chinese medicine database for drug screening in silico. PLoS ONE 6(1):e15939. doi: https://doi.org/10.1371/journal.pone.0015939

11. Ru J, Li P, Wang J, Zhou W, Li B, Huang C et al (2014) TCMSP: a database of systems pharmacology for drug discovery from herbal medicines. J Cheminform 6:13. doi: https://doi.org/10.1186/1758-2946-6-13
12. Huang L, Xie D, Yu Y, Liu H, Shi Y, Shi T et al (2018) TCMID 2.0: a comprehensive resource for TCM. Nucleic Acids Res 46(D1):D1117–D20. doi: https://doi.org/10.1093/nar/gkx1028
13. Ye H, Ye L, Kang H, Zhang D, Tao L, Tang K et al (2011) HIT: linking herbal active ingredients to targets. Nucleic Acids Res 39(Database issue):D1055–D1059. doi: https://doi.org/10.1093/nar/gkq1165
14. Dong J, Wang NN, Yao ZJ, Zhang L, Cheng Y, Ouyang D et al (2018) ADMETlab: a platform for systematic ADMET evaluation based on a comprehensively collected ADMET database. J Cheminform 10(1):29. doi: https://doi.org/10.1186/s13321-018-0283-x
15. Xiong GL, Wu ZX, Yi JC, Fu L, Yang ZJ, Hsieh CY et al (2021) ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties. Nucleic Acids Res 49(W1):W5–W14. doi: https://doi.org/10.1093/nar/gkab255
16. Wu Y, Zhang F, Yang K, Fang S, Bu D, Li H et al (2019) SymMap: an integrative database of traditional chinese medicine enhanced by symptom mapping. Nucleic Acids Res 47(D1):D1110–D7. doi: https://doi.org/10.1093/nar/gky1021
17. Kim S, Chen J, Cheng T, Gindulyte A, He J, He S et al (2021) PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 49(D1):D1388–D95. doi: https://doi.org/10.1093/nar/gkaa971
18. Kim S, Thiessen PA, Bolton EE, Bryant SH (2015) PUG-SOAP and PUG-REST: web services for programmatic access to chemical information in pubchem. Nucleic Acids Res 43(W1):W605–W611. https://doi.org/10.1093/nar/gkv396
19. Wang NN, Dong J, Deng YH, Zhu MF, Wen M, Yao ZJ et al (2016) ADME Properties evaluation in drug discovery: prediction of Caco-2 cell permeability using a combination of NSGA-II and boosting. J Chem Inf Model 56(4):763–773. https://doi.org/10.1021/acs.jcim.5b00642
20. Wang NN, Deng ZK, Huang C, Dong J, Zhu MF, Yao ZJ et al (2017) ADME properties evaluation in drug discovery: prediction of plasma protein binding using NSGA-II combining PLS and consensus modeling. Chemometr Intell Lab Syst 170:84–95. doi: https://doi.org/10.1016/j.chemolab.2017.09.005
21. Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, Owen G et al (2016) ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminform 8:61. doi: https://doi.org/10.1186/s13321-016-0174-y
22. Gadaleta D, Lombardo A, Toma C, Benfenati E (2019) Correction to: a new semi-automated workflow for chemical data retrieval and quality checking for modeling applications. J Cheminform 11(1):31. doi: https://doi.org/10.1186/s13321-019-0353-8
23. Keiser MJ, Roth BL, Armbruster BN, Ernsberger P, Irwin JJ, Shoichet BK (2007) Relating protein pharmacology by ligand chemistry. Nat Biotechnol 25(2):197–206. doi: https://doi.org/10.1038/nbt1284
24. Daina A, Michielin O, Zoete V (2019) SwissTargetPrediction: updated data and new features for efficient prediction of protein targets of small molecules. Nucleic Acids Res 47(W1):W357–W64. doi: https://doi.org/10.1093/nar/gkz382
25. Hamad S, Adornetto G, Naveja JJ, Chavan Ravindranath A, Raffler J, Campillos M (2019) HitPickV2: a web server to predict targets of chemical compounds. Bioinformatics 35(7):1239–1240. doi: https://doi.org/10.1093/bioinformatics/bty759
26. Awale M, Reymond JL (2017) The polypharmacology browser: a web-based multi-fingerprint target prediction tool using ChEMBL bioactivity data. J Cheminform 9:11. doi: https://doi.org/10.1186/s13321-017-0199-x
27. Awale M, Reymond JL (2019) Polypharmacology browser PPB2: target prediction combining nearest neighbors with machine learning. J Chem Inf Model 59(1):10–17. https://doi.org/10.1021/acs.jcim.8b00524
28. Davies M, Nowotka M, Papadatos G, Dedman N, Gaulton A, Atkinson F et al (2015) ChEMBL web services: streamlining access to drug discovery data and utilities. Nucleic Acids Res 43(W1):W612–W620. doi: https://doi.org/10.1093/nar/gkv352
29. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Felix E et al (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47(D1):D930–D40. doi: https://doi.org/10.1093/nar/gky1075
30. UniProt C (2021) UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49(D1):D480–D9. doi: https://doi.org/10.1093/nar/gkaa1100
31. Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A et al (2018) Unexplored therapeutic opportunities in the human genome. Nat Rev Drug Discov 17(5):317–332. doi: https://doi.org/10.1038/nrd.2018.14
32. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR et al (2018) DrugBank 5.0: a major update to the drugbank database for 2018. Nucleic Acids Res 46(D1):D1074–D82. https://doi.org/10.1093/nar/gkx1037
33. Skinnider MA, Dejong CA, Franczak BC, McNicholas PD, Magarvey NA (2017) Comparative analysis of chemical similarity methods for modular natural products with a hypothetical structure enumeration algorithm. J Cheminform 9(1):46. doi: https://doi.org/10.1186/s13321-017-0234-y
34. Vogt M, Stumpfe D, Geppert H, Bajorath J (2010) Scaffold hopping using two-dimensional fingerprints: true potential, black magic, or a hopeless endeavor? guidelines for virtual screening. J Med Chem 53(15):5707–5715. https://doi.org/10.1021/jm100492z
35. Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T et al (2009) KNIME - the konstanz information miner: version 2.0 and beyond. SIGKDD Explor Newsl 11(1):26–31. https://doi.org/10.1145/1656274.1656280
36. Zhao HM, Zhang XY, Lu XY, Yu SR, Wang X, Zou Y et al (2018) Erzhi pill(®) protected Experimental Liver Injury against apoptosis via the PI3K/Akt/Raptor/Rictor pathway. Front Pharmacol 9:283. https://doi.org/10.3389/fphar.2018.00283
37. Pang Z, Zhi-yan Z, Wang W, Ma Y, Feng-ju N, Zhang X et al (2015) The advances in research on the pharmacological effects of fructus ligustri lucidi. Biomed Res Int. https://doi.org/10.1155/2015/281873
38. Pan B, Pan W, Lu Z, Xia C (2021) Pharmacological mechanisms underlying the hepatoprotective effects of ecliptae herba on hepatocellular carcinoma. Evid Based Complement Alternat Med 2021:5591402. https://doi.org/10.1155/2021/5591402
39. Xu P, Su S, Tan C, Lai RS, Min ZS (2017) Effects of aqueous extracts of Ecliptae herba, polygoni multiflori radix praeparata and rehmanniae radix praeparata on melanogenesis and the migration of human melanocytes. J Ethnopharmacol 195:89–95. https://doi.org/10.1016/j.jep.2016.11.045
40. Gao L, Li C, Wang Z, Liu X, You Y, Wei H et al (2015) Ligustri lucidi fructus as a traditional chinese medicine: a review of its phytochemistry and pharmacology. Nat Prod Res 29(6):493–510. doi: https://doi.org/10.1080/14786419.2014.954114
41. Huang da W, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1):44–57. doi: https://doi.org/10.1038/nprot.2008.211
42. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K (2017) KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45(D1):D353–D61. doi: https://doi.org/10.1093/nar/gkw1092
43. Quinones-Lombrana A, Li N, Del Solar V, Atilla-Gokcumen GE, Blanco JG (2018) CBR1 rs9024 genotype status impacts the bioactivation of loxoprofen in human liver. Biopharm Drug Dispos 39(6):315–318. doi: https://doi.org/10.1002/bdd.2135
44. Garcia-Ruiz I, Blanes Ruiz N, Rada P, Pardo V, Ruiz L, Blas-Garcia A et al (2019) Protein tyrosine phosphatase 1b deficiency protects against hepatic fibrosis by modulating nadph oxidases. Redox Biol 26:101263. doi: https://doi.org/10.1016/j.redox.2019.101263
45. Zhou HZ, Zeng HQ, Yuan D, Ren JH, Cheng ST, Yu HB et al (2019) NQO1 potentiates apoptosis evasion and upregulates XIAP via inhibiting proteasome-mediated degradation SIRT6 in hepatocellular carcinoma. Cell Commun Signal 17(1):168. doi: https://doi.org/10.1186/s12964-019-0491-7
46. Chu B, Kon N, Chen D, Li T, Liu T, Jiang L et al (2019) ALOX12 is required for p53-mediated tumour suppression through a distinct ferroptosis pathway. Nat Cell Biol 21(5):579–591. doi: https://doi.org/10.1038/s41556-019-0305-6
47. Chen J, Liu QM, Du PC, Ning D, Mo J, Zhu HD et al (2020) Type-2 11beta-hydroxysteroid dehydrogenase promotes the metastasis of colorectal cancer via the Fgfbp1-AKT pathway. Am J Cancer Res 10(2):662–673

48. Xiao X, Wang W, Li Y, Yang D, Li X, Shen C et al (2018) HSP90AA1-mediated autophagy promotes drug resistance in osteosarcoma. J Exp Clin Cancer Res 37(1):201. doi: https://doi.org/10.1186/s13046-018-0880-6
49. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D et al (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13(11):2498–2504. doi: https://doi.org/10.1101/gr.1239303

Publisher's Note