Faculty of Biological and Environmental
Sciences
University of Helsinki
Finland

Network Pharmacology Approaches for Understanding Traditional Chinese Medicine

Yinyin Wang

Research Program in Systems Oncology
Institute for Molecular Medicine Finland
(FIMM)
And
Doctoral Programme in Integrative Life Science
University of Helsinki

ACADEMIC DISSERTATION

To be presented for public examination with permission of the
Faculty of Biological and Environmental Sciences of the
University of Helsinki, in zoom, on the 8th of November 2021,
at 3:15 pm.

Helsinki 2021

Supervised by

Asst. Prof. Jing Tang, PhD
Research Program in Systems Oncology
Faculty of Medicine
University of Helsinki
Helsinki, Finland

Mohieddin Jafari, PhD
Research Program in Systems Oncology
Faculty of Medicine
University of Helsinki
Helsinki, Finland

Thesis advisory committee

Henri Xhaard, PhD
Faculty of Pharmacy
University of Helsinki
Helsinki, Finland

Jianwei Li, PhD
Department of Chemistry
University of Turku
Turku, Finland

Thesis reviewers

Prof. Aik Choon Tan, PhD
Department of Biostatistics and Bioinformatics
H. Lee Moffitt Cancer Center and Research Institute
Tampa, FL, USA

Prof. Feng Zhu, PhD
College of Pharmaceutical Sciences
Zhejiang University
Hangzhou, China

Opponent

Asst. Prof. Feixiong Cheng, PhD
Genomic Medicine Institute
Cleveland Clinic
Cleveland, US

Custos

Prof. Kari P Keinänen, PhD
Faculty of Biological and Environmental Sciences
University of Helsinki
Helsinki, Finland

ISBN 978-951-51-7539-7 (paperback)
ISBN 978-951-51-7540-3 (PDF)
https://ethesis.helsinki.fi/
Unigrafia
Helsinki 2021
The Faculty of Biological and Environmental Sciences uses the Urkund system (plagiarism recognition) to examine all doctoral dissertations.

"*We are what we repeatedly do; excellence, then, is not an act but a habit.*"

Aristotle

TABLE OF CONTENTS

LIST OF ORIGINAL PUBLICATIONS

This thesis is based on the following publications, which are referred to in the text by their Roman numerals:

I. **Wang Y**, Jafari M, Tang Y, Tang J. Predicting Meridian in Chinese traditional medicine using machine learning approaches. ***PLoS Comput Biol***. 2019;15(11):e1007249.

II. **Wang Y**, Yang H, Chen L, Jafari M, Tang J. Network-based Modeling of Herb Combinations in Traditional Chinese Medicine. ***Brief Bioinform***, **2021; bbab106**

III. Jafari M, **Wang Y**, Amiryousefi A, Tang J. Unsupervised Learning and Multipartite Network Models: A Promising Approach for Understanding Traditional Medicine. ***Front Pharmacol***. 2020;11:1319.

The following publications are related to the study but are not included in the thesis:

- Liu M*, **Wang Y***, Miettinen JJ, Kumari R, Majumder MM, Tierney C, …, Tang J, Heckman A, S100 calcium binding protein family members associate with poor patient outcome and response to proteasome inhibition in multiple myeloma. **Front Cell Dev Biol.** 2021. 9: p. 723016.
- Zagidullin B, Aldahdooh J, Zheng S, Wang W, **Wang Y**, Saad J, *et al*. DrugComb: An integrative cancer drug combination data portal. ***Nucleic Acids Res***. 2019;47(W1):W43-w51.
- Hsieh K, **Wang Y**, Chen L, Zhao Z, Savitz S, Jiang X, *et al*. Drug Repurposing for COVID-19 using Graph Neural Network with Genetic, Mechanistic, and Epidemiological Validation**. ArXiv**. 2020.

ABBREVIATIONS

ADME	Absorption, distribution, metabolism, excretion, and toxicity
AI	Artificial intelligence
AIDS	Deficiency syndrome
ANNs	Artificial neural networks
AUC	Area under curve
AUPRC	Area under curve (AUC) of precision and recall (PR)
BMI	Body Mass Index
BNs	Bayesian Networks
CDW	Clinical data warehouse
cGAP	Good Agricultural Practices
cGMP	Good Manufacturing Practices
DEGs	Differential expressed genes
DisGeNET	Database of gene–disease associations
DL	Drug-likeness
DMIM	Distance-Based Mutual Information Model
DT	Decision tree
Ext	Extended fingerprint
FDR	False positive rate
FN	False negative
FP	False positive
GO	Gene Ontology
GEO	Gene Expression Omnibus Database
H1N1	Anti-influenza A
ISN	Ingredient similarity network
kNN	k-nearest neighbour
KEGG	kyoto encyclopedia of genes and genomes
LDA	linear discriminant analysis
LoR	Logistic regression
LR	Linear least-squares regression

LWDH	Liu-Wei-Di-Huang
MACCS	Molecular Access System
MCC	Matthews correlation coefficient
ML	Machine learning
MM	Modern Medicine
MOAs	Mechanisms of Action
MR	Molar refractivity
NIH	Heterogeneous information network
NPs	Natural products
NSN	Natural product similarity network
OB	Oral bioavailability
OMIM	Online Mendelian Inheritance in Man
PDB	Protein Data Bank
PDD	Phenotypic drug discovery
PNN	Probabilistic neural network
PPI	Protein–protein network
PR	Precision Recall
RF	Random forest
ROAUC	Receiver operating character characteristic (ROC) curve (AUC)
ROC	Receiver operating character characteristics
SEA	Similarity ensemble approach
SOM	Self-organizing map
Sub	Substructure fingerprint
SVM	support vector machine
TCM	Traditional Chinese Medicine
TDD	Target-based drug discovery
TM	Traditional medicine
TN	True negative
TP	True positive
TPP	Thermal proteomics profiling
TPR	True positive rate
TTD	Therapeutic Target Database

UTCMLS Unified traditional Chinese medical language system

ABSTRACT

Traditional Chinese medicine (TCM) has obvious efficacy on disease treatments and is a valuable source for novel drug discovery. However, the underlying mechanism of the pharmacological effects of TCM remains unknown because TCM is a complex system with multiple herbs and ingredients coming together as a prescription. As evidence-based science, there are many special theories in the TCM system that do not exist in modern medicine. For instance, TCM treats the human body as a holistic system in a balance of YIN and YANG. Another example is the meridian classification of herbs in TCM based on treatment selection for diseases in different positions of the human body. Some of these theories have been validated at the molecular level with the development of the system biological concepts. However, a lack of understanding of these theories is still the main bottleneck of their wide application. Therefore, it is urgent to apply computational tools to TCM to understand the underlying mechanism of TCM theories at the molecular level. TCM emphasizes a specified herb combination as a prescription based on the symptoms of the individual, which is also a kind of precision medicine at the biological level. Hence, it is also meaningful to use advanced network algorithms to explore potential effective ingredients and illustrate the principles of TCM in system biological aspects.

In this thesis, we aim to understand the underlying mechanism of actions in complex TCM systems at the molecular level by bioinformatics and computational tools. In study **I**, a machine learning framework was developed to predict the meridians of the herbs and ingredients. In detail, parameter optimization, model selection and feature selection were performed on four supervised classification methods (SVM, DT, RF and kNN) with features from four different types of fingerprints (Substructure, PubChem, MACCS, Extended fingerprint) and absorption, distribution, metabolism, excretion and toxicity (ADME) properties. Finally, we achieved high accuracy of the meridians prediction for herbs and ingredients, suggesting an association between meridians and the molecular features of ingredients and herbs, especially the most important features for machine learning models.

Secondly, we proposed a novel network approach to study the TCM formulae by quantifying the degree of interactions of pairwise herb pairs in study **II** using five network distance methods,

including the closest, shortest, central, kernel, as well as separation. We demonstrated that the distance of top herb pairs is shorter than that of random herb pairs, suggesting a strong interaction in the human interactome. In addition, center methods at the ingredient level outperformed the other methods. It hints to us that the central ingredients play an important role in the herbs. Our network modelling provides a novel systems medicine framework to characterise herb interactions and may further suggest novel herb combinations and potential ingredient combinations based on their synergistic interactions.

Thirdly, we explored the associations between herbs or ingredients and their important biological characteristics in study **III**, such as properties, meridians, structures, or targets via clusters from community analysis of the multipartite network. We found that herbal medicines among the same clusters tend to be more similar in the properties, meridians. Similarly, ingredients from the same cluster are more similar in structure and protein target. These findings might provide novel insight for the understanding of TCM, suggesting that multipartite network models from TCM are suitable to study TCM from system biological aspects.

In summary, this thesis intends to build a bridge between the TCM system and modern medicinal systems using computational tools, including the machine learning model for meridian theory, network modelling for TCM formulae, as well as multipartite network analysis for herbal medicines and their ingredients. We demonstrated that applying novel computational approaches on the integrated high-throughput omics would provide insights for TCM and accelerate the novel drug discovery as well as repurposing from TCM.

INTRODUCTION

Modern medicine systems have been developing rapidly since the discovery of penicillin in 1928. In modern medicine, the one-drug–one-target–one-disease drug discovery strategy has been used for many years with considerable success. However, many limitations and obstacles of this strategy have emerged gradually, including the lack of efficacy, drug resistance, and side effects, especially in complex diseases, such as cancer [1] and diabetes [2]. As a result, the Polypharmacology concept has been proposed and is becoming popular as a paradigm shift of drug discovery [3-10]. Polypharmacology concept means using multiple drugs on multiple targets to treat disease as an entire system in the format of multi-drug combination. Whereas, it remains a daunting task to search for all the potential synergistic drug combinations by wet experiments [11, 12]. Thus, it is of immense importance to develop cost-effective computational tools for searching potential drug combinations. Fortunately, in East Asia, traditional medicine has been using plant-derived natural products in the format of herb combination, also known as herb formulae, for disease prevention and treatment for thousands of years, especially in TCM [13, 14]. Herb formulae often combine a few herbs with multiple bioactive ingredients and thus produce synergistic effects in a personalized medicine manner. In this way, herb formulae perform with maximal therapeutic efficacy and minimal side effects[15].

Notably, many formulae have been on the market as capsules or injections for disease treatment across the world [16]. These formulae, from thousands of years of clinical experience, are among the valuable resources for drug combination discovery. Despite the obvious therapeutic effects on patients, the underlying synergistic mechanism remains unknown to us mainly because the TCM prescription is carried out under the guidelines of the principles of TCM theories.

Traditional medicine (TM) treats the human body as a miniature universe and can be classified as five interacting elements: metal, wood, water, fire, and earth [17]. In addition, in TM, it was believed that the disease condition of the human body and corresponding symptoms are caused by the loss of balance [18, 19]. Thus, the process of disease treatment is to restore balance in the system [20] with either physical methods, such as acupuncture [21, 22], or chemical

approaches, such as herbal formulae. More importantly, there are precise medical concepts in TCM that classify diseases by Yin and Yang or certain inner channels of the body known as meridians [23]. In addition, according to organ selective distribution, herbs can be assigned to 12 principal meridians. However, the meridian concept in TCM refers to the whole system of functionally related organs rather than the physical organ definition used in modern medicine. The meridian theory has existed and developed for thousands of years as one of the fundamental theories to guide TCM practice, whereas the underlying molecular basis of meridian classification for herbs is yet to be uncovered. Hence, to understand the complex system like TCM, it is crucial to explore the fundamental theories developed from the evidence of TCM practice on patients [24, 25].

Researchers have attempted to explain meridian with fascia networks [26] and perivascular space [27], but neither of them were confirmed by experiments. Although there is no widely accepted anatomical or physiological evidence for meridians, the meridian concept has been practiced for thousands of years in TCM to guide the prescription in the clinic [27-29]. Although the rationale of meridians for specific herbs has been investigated, such as *Platycodi Radix* (Jiegeng) [30] and *Salvia miltiorrhiza burge* (Danshen) [31, 32], there has been limited larger-scale analysis on meridians and other principles at the molecular level in TCM [33]. In fact, the biochemical and pharmacological properties of the bioactive ingredients in TCM herbs have been gradually uncovered by advanced technologies, which make the study at the molecular level possible.

Recently, machine learning techniques have been utilized in the study of herbal formulae or drug combinations [34-36], drug–target networks of natural products [37], and the hot or cold nature of the herbs [38, 39]. However, machine learning studies focused on meridian classification by the chemical structure and physicochemical features of ingredient compounds are rare.

Returning to the combination of herbs, computational methods have been developed to investigate the properties of herb pairs of herbal formulae, suggesting the feasibility of using machine learning and network modelling to understand herb–herb interactions [40-43]. Herb pairs in herbal formulae refer to a unique combination of two specific herbs, which are the most fundamental elements of TCM formulae [13]. For different diseases, various herbs with

different biological functions can be added into basic herb pairs as a more complex formula [40, 41]. Therefore, it is desirable to illustrate the rationale of certain herb pairs for a particular disease [33, 44]. However, one of the major bottlenecks is that a herb combination is inherently more complex than a drug combination, as herbs usually consist of multiple ingredients. Due to the multiple-ingredient system of herbs, it is difficult to determine the ingredients with synergistic effects and the underlying mechanism of actions (MOAs) of the therapeutic effects [43, 45-47]. Recent studies have elucidated that the synergistic effects of herb combinations might be associated with the interactions between their ingredients [48-51] at protein levels. Based on the remarkable success of network-based models in understanding the interaction between chemicals and diseases [52-54] in modern drug combinations, we can hypothesize that network pharmacology models based on the ingredient–target interactions would be promising to explore the herb pairs' MOAs and further facilitate the phenotypic-based drug discovery from existing TCM herbs or ingredients.

Furthermore, classification and clustering approaches are widely applied to elucidate the hidden relationships in complex systems, such as human biology and diseases. In fact, modularity analysis based on networks with nodes and edges is one special kind of classification to detect the subnetwork with specific characteristics [55, 56]. In detail, the nodes in a community detected from a network analysis would have a high intra-relationship and exhibit low inter-relationship with the other nodes outside the community [57]. Thus, it is meaningful to perform network module analysis in order to uncover the complex topology in networks. To understand the complex biological process systematically, such as signalling pathways or target-target interactions, it is of great importance to construct multipartite networks [58]. A multipartite network is a network that consists of multiple sets of nodes and edges by integrating mixed types of datasets from uni-partite networks with single sets of nodes and edges, which make it promising to detect complex underlying associations in biology [59].

Our study first investigated the classification of meridians by machine learning methods and revealed that meridians of herbs is predictable by machine learning methods with the molecular fingerprints and ADME properties as features. In addition, we demonstrated that the meridian concept has the molecular basis, which helps to understand the meridians at the molecular level. Furthermore, network-based models were built to quantify the interaction distance of a herb

pair within a protein–protein interaction network at the herbal, ingredient, and protein target levels. We adopted five distance metrics, including the closest, shortest, separate, kernel and center. By comparing the distances of random herb pairs and top herb pairs, we demonstrated that the underlying mechanism of herb combination is also derived from affecting neighbouring proteins in the human interactome. More importantly, the central active ingredients from individual herbs play an important role in herb combination. Also, a modularity analysis of multipartite networks was adopted to investigate the associations between TCM classifications and the chemical or biological features of the herbs or ingredients. To conclude, this thesis developed novel computational approaches to understand TCM systematically, which may help to identify the MOAs or active compounds of TCM theories or therapeutic effects and promote the modernization of TCM.

1. REVIEW OF LITERATURE

1.1. Natural products (NPs)

1.1.1. Vital role of NPs in drug discovery

Natural products refer to a series of chemical compounds or substances produced by living organisms (cells, tissues, and secretions of microorganisms, plants, and animals) through the pathways of primary or secondary metabolism [60]. Secondary metabolites are significant for the medical field, as these metabolites are not only essential for survival itself but also produce various pharmacological activities in disease treatment because of their chemical diversity in nature [61, 62]. As a result, for many years, researchers have put effort into the detection, isolation and identification of bioactive compounds, which is called bioprospecting [63]. However, these processes are relatively time-consuming and expensive on a commercial scale, which makes large-scale screening from natural products (NPs) challenging. Thus, producing the new compound by total synthesis or semi-synthesis based on the known structure of natural products was proposed.

1.1.2. Drugs from NPs

Over these years, with the development of computer science, computational approaches have become powerful tools for drug screening, design and modification, which further promotes the boost of natural products in drug discovery [64]. In addition, some virtual and physical natural product libraries or databases were constructed to help the virtual screening [65]. Consequently, numerous drug discoveries used NPs as a starting point and have achieved considerable success, especially analgesics, anti-infectives, anti-tumour drugs and HMG-CoA reductase inhibitors [66].

A well-known example is finding salicylic acid from the bark of the willow tree. The bark of the willow tree has been well known for its pain-relieving effects for a long history. To detect compounds with such bioactivity, compounds in the bark of the willow tree were isolated and tested. Finally, it turned out that the compound salicin can act against pain after it is hydrolyzed into salicylic acid by inhibiting the cyclooxygenase (COX) enzyme [67]. Furthermore, a

considerable number of anti-infection drugs, such as penicillin from Penicillium, are based on NPs [63]. Products have also been widely used to treat hypertension as well as congestive heart failure.

1.1.3. Relationship between NPs and traditional medicine

Traditional medicine is also known as indigenous or folk medicine. According to the official definition from the World Health Organization (WHO), traditional medicine is "the sum total of the knowledge, skills, and practices based on the theories, beliefs, and experiences indigenous to diverse cultures, whether explicable or not, used in the maintenance of health as well as in the prevention, diagnosis, improvement or treatment of physical and mental illness." [68]

Traditional medicine is evidence-based and has played a key role since ancient times in disease prevention and treatment. With thousands of years' practice, traditional medicine has developed into various branches in different regions [54]. For instance, one of the main branches is traditional Chinese medicine from China. Besides traditional Chinese medicine, there are various kinds of traditional medicine, such as Traditional European medicine from Europe, as well as traditional indigenous medicine from Assam and the rest of NE India, Siddha medicine from South India, traditional Korean medicine from Korea, and Ayurveda from Indian. In addition, traditional African medicine referred to a range of traditional medicine disciplines involving indigenous herbalism and African spirituality). Unani is the perso-Arabic traditional medicine practiced in Muslim cultures in South Asia and modern day Central Asia while Islamic medicine is the science of medicine developed in the Middle East.

However, many of these traditional medicines are claimed to be pseudoscience because the claimed effects or concepts in some traditional medicines are incompatible with the scientific facts or cannot be verified by scientific methods [69]. With the developments of the chemical field, the advanced methods of compound extraction, isolations and purification accelerate drug discovery by extracting pure natural products with biological activity, which led to modern medicine (MM) [70]. In modern medicine, traditional medicine is considered as alternative medicine, which refers to any practice that tends to achieve the healing effects of medicine but lacks biological plausibility and is untested, or untestable [71]. However, considering the

considerable success of natural products in drug discovery and some drawbacks of MM, it is time to reconsider the role of traditional medicine and exploit it with advanced technologies and approaches.

1.2. Traditional Chinese medicine (TCM)

1.2.1. Developing history of TCM

As an important branch of traditional medicine, the history of traditional Chinese medicine goes back to 1100 Before Christ (B.C.). The ancient Chinese ancient book "Wu Shi Er Bing Fang" was written at that time and recorded 52 prescriptions. Additionally, 365 Chinese medicines were recorded in the book "Shennong Herbal" (~100 B.C.), and 850 herbal medicines are found in the book named "Tang Herbal" from 659 After Christ [72].

TCM not only played an important role in disease treatment and prevention in ancient times but also is of great significance for the development of modern medicine, especially for novel drug discovery. For example, more than 8,000 TCM components have been reported to have various pharmacological effects. Particularly, TCM has become increasingly popular in the drug discovery field in recent years and there are a larger number of clinical TCM-related trials in clinicaltrials.gov. A good example is the discovery of artemisinin. In 1972, Chinese scientists discovered that the compound artemisinin derived from the herb *Artemisia annua* (青蒿, qinghao) has antimalarial effects [73]. Another example is the inorganic compound arsenic trioxide (As2O3), which shows an obvious therapeutic effect for acute promyelocytic leukemia [74]. This compound has been used in TCM for a long time to cure erosion sore rot and malaria.

Despite these successful applications of TCM, there is a giant gap between TCM systems and modern medical systems, as these two systems are under different theoretical guidelines. For instance, the syndrome and symptom classification used in TCM are different from the symptoms we use in modern medicine. These theories go beyond immediately perceivable sensations, although they have been proved effective in clinical practice. Therefore, it is of immense importance to understand the underlying molecular basis of these theories using computational tools so that we can build up a bridge between TCM and modern medical systems.

1.2.2. Theories in TCM

In TCM, many theories were developed from thousands of years of clinical experience for disease treatment. Some theories are related to the differentiation of the human body, such as meridian Tropism theory, Yin-Yang theory, and Five elements theory. At the same time, some theories are used to characteristic medicines properties, e.g. Siqi theory, Wuwei theory, and Jun Chen Zuo Shi theory [38]. For instance, herbs are classified according to their flavours (sour, salty, sweet, bitter, and pungent), which is called Wuwei theory
Yin-Yang theory [75] is used to describe the physiological and pathological states of the human body. TCM believes that the imbalance between Yin and Yang is the cause of disease. Consequently, the goal of disease prevention and treatment are to restore the body's Yin and Yang to a balanced state.

Siqi theory [76] is defined as four main reactions of the human body (cold, hot, warm, or cool) after the administration of a specific medicine. For example, chewing a mint (*Mentha spicata*) leaf elicits a cold feeling, while masticating a piece of ginger (*Zingiber officinale*) leads to a hot sensation. A neutral remedy, such as Wolfberry (*Lycium barbarum*), may not be associated with an obvious cold or hot feeling.

Meridian tropisms referred to 14 meridian channels (main channels) distributed longitudinally throughout the human body [77]. These meridians are related to disease-associated organs but are not the exact same as organs in modern medicine [78]. For example, *Platycodon grandiflorum* belongs to the lung meridian in TCM. In fact, it was reported that the active ingredient saponin in *Platycodon grandiflorum* has a close functional relationship with the lung and respiratory systems [30]. Furthermore, *Salvia miltiorrhiza burge* is classified as both Heart meridian and Liver meridian and thus this herb can be used to treat cardiovascular diseases and hepatitis [31] accordingly.

1.2.3. TCM syndromes and precision medicine

At present, the classification of diseases in precision medicine relies on DNA abnormalities, such as gene mutation, over-amplification, large-deletions, or recombination [79]. Thus, one of the important strategies for precision medicine is to identify the DNA abnormality and

discover drugs targeted on these abnormalities. However, there are some diseases not caused by genetic sequence abnormalities, especially those complex diseases. In contrast, the human body reflects individual differences, with different epigenetic pressures. Interestingly, there is a report that 29.8% of the 3294 TCM medicinals affect the epigenomes and miRNA expression of human cells [80]. In fact, there has been a long history of precision medicine in the TCM system in the processes of diagnosis and prescription.

The syndrome is an important concept to realize precision medicine in TCM. The syndrome is a group of specific symptoms of diseases in the human body, which is also known as Pattern, or co-module, or Zheng [81]. TCM clinical practitioners would make diagnoses based on TCM syndromes of individual patients and give the corresponding herbs to adjust the disorders in the human body.

One interesting rule in TCM is “the same treatment for different diseases” and “the same disease with different treatments.” This reflects TCM practitioners’ deep thoughts about the formula–syndrome relationship. It was reported that herbal formulae and TCM Syndrome can be explained well by co-module analysis [82]. Liu-Wei-Di-Huang (LWDH) formula is a good example of “the same treatment for different diseases” as LWDH has been used to treat multiple diseases—cancer, diabetes, and hypertension. In addition, it was found that the key genes in the network of LWDH formula are significantly close to the genes associated with these diseases and these LWDH-treated diseases share an overlapping molecular basis and show high phenotypic similarity. Therefore, the same LWDH formula can be used to treat different diseases[15].

Arthritis is a good example of “the same disease with different treatments”. The Cold Syndrome and Hot Syndrome are the two most common and representative syndrome classes that represent two opposite conditions. For example, in the arthritis rat model, the hot formula (Wen-Luo-Yin formula) is more effective on the hub nodes of the Cold Syndrome network, whereas the cold formula (Qing-Luo-Yin formula) tends to target the hub nodes of the Hot Syndrome network [83, 84]. This indicates that the same disease with different syndromes (cold and hot) should be treated with different formulae.

1.3. The data sources for TCM research

The rapid development of molecular profiling technologies for extraction and identification helps research uncover the ingredients in herbal constituents [85]. In addition, with the rapidly emerging new omics technologies, people have advanced research tools to study complex problems, especially in the TCM area. Integration of these various kinds of omics data is of great significance for understanding MOAs at multiple levels, such as cells, tissues, organs, and organisms, as well as disease levels, which further leads to precision medicine. By detecting the molecular change in the human body (expression or protein or metabolites, etc.) after treatments with TCM, we can learn more about the precision medicine strategy of TCM. With the development of TCM pharmacology, an increasing number of databases of TCM were created [86].

Most of the latest databases of TCM provide information about prescriptions, herbs, ingredients, targets, diseases and their interaction relationships. These databases also enable users to perform network pharmacology analysis. However, the entire workflow for TCM is not as mature as its application in modern drugs although the omics data was widely applied in TCM, especially in system biology, and there are also a larger number of advanced methods can be utilized to make full use of omics [87].

1.3.1. TCM databases

TCM-ID

For instance, TCM-ID [88] (http://bidd.group/TCMID/, Traditional Chinese Medicine Information Database) was developed in 2006 to provide comprehensive information on TCM, including prescriptions (n=1,588), their constituent herbs (n=1,313 herbs), herbal ingredients (n=5,669), and corresponding molecular information (n=3,725).

Database@taiwan

Database@taiwan [89] (http://tcm.cmu.edu.tw/), developed in 2011, was the world's largest non-commercial TCM database at the time, including 20,000 pure compounds and 435 TCM herbs. The number of compounds in this database increased by about 61,000 in 2013.

TCMSP

TCMSP [90] (Traditional Chinese Medicine Systems Pharmacology Database, https://tcmspw.com/tcmsp.php) was published in 2012 and updated in 2014, including 499 Chinese herbs covering 29,384 ingredients, 3,311 targets and 837 associated diseases. In this database, a comprehensive network between Herbs–Compounds–Targets–Diseases (H–C–T–D) was created. TCMSP also systematically calculated the 12 ADME properties for the first time to filter the ingredients in TCM that have poor oral absorbability and low drug-likeness.

TCMID

The Traditional Chinese Medicine Integrated Database (TCMID, http://119.3.41.228:8000/tcmid/) [91] integrates the data from Database@Taiwan, and Herb Ingredients' Targets (HIT) [92] (http://lifecenter.sgst.cn/hit/) and was updated in 2018, including 49,000 prescriptions, 8,159 herbs, 25,210 ingredients, 3,791 diseases, 6,828 drugs, and 17,521 targets.

TCM-Mesh

TCM-Mesh [93] (http://mesh.tcm.microbioinformatics.org/#) was published in 2017, including 6,235 herbs, 383,840 compounds, 14,298 genes, 6,204 diseases, 144,723 gene–disease associations and 3,440,231 pairs of gene interactions. Moreover, there are 34,144 herb–ingredient pairs, covering 10,140 ingredients and 6,235 herbs. TCM-Mesh also provides 163,221 side effect records (1,430 ingredients and 6,123 side effects) from TOXNET, and 71 toxic records for 71 ingredients from SIDER [94].

YaTCM

YaTCM [95] (Yet another Traditional Chinese Medicine database, http://cadd.pharmacy.nankai.edu.cn/yatcm/home) is a free web-based toolkit published in 2018. It contains 47,696 natural compounds, 6,220 herbs, 18,697 targets (with 3,461 therapeutic targets included), 1,907 predicted targets, 390 pathways, and 1,813 prescriptions with manually curated information.

ETCM

ETCM [96] (The Encyclopedia of Traditional Chinese Medicine, http://www.tcmip.cn/ETCM/index.php/Home/Index/) is another web server tool established for the network analysis of TCM in 2018, including herbs (402), formulae (3,959) and ingredients (7,284). More importantly, ETCM not only includes the indication but also the Administration, Syndromes in Chinese, Syndromes in English and Dosage Form. For herbs, many quality-control characteristics were curated in ETCM.

TCMAnalyzer

TCMAnalyzer [97] (A Chemo- and Bioinformatics Web Service for Analyzing Traditional Chinese Medicine, http://www.rcdd.org.cn/tcmanalyzer) supports substructure search, similarity search, and scaffold search from the aspects of chemoinformatics.

BATMAN-TCM

BATMAN-TCM [98] (a Bioinformatics Analysis Tool for Molecular mechANism of Traditional Chinese Medicine, http://bionet.ncpsb.org.cn/batman-tcm/) is an online bioinformatics analysis tool that supports user-customized input and interactive analysis of the molecular mechanism of combinations of formulae or herbs in the holistic aspects.

TM-MC

TM-MC [99] (http://informatics.kiom.re.kr/compound/) is a database about Northeast Asian traditional medicine. TM-MC extracted 14,000 chemical compounds from 536 medicinal materials and 4,000 journal articles in MEDLINE and PMC.

SymMap

SymMap [100] (symptom mapping, https://www.symmap.org/) is an integrative database of traditional Chinese medicine that maps symptoms in TCM to modern symptoms and further diseases, covering 1,717 TCM symptoms, 499 herbs, 961 modern symptoms, 5,235 modern diseases, 4,302 targets, and 19,595 ingredients. SymMap completes the bridge from TCM

symptoms and diseases to modern symptoms and diseases at both the phenotypic and molecular levels.

HERB

The HERB [101] (high-throughput experiment and reference-guided database of traditional Chinese medicine, http://herb.ac.cn/) database was constructed in 2020. HERB not only provides gene expression profiling but also manually collects high-confidence targets and diseases from the literature for ingredients and herbs in TCM. In detail, this database contains 6,164 gene expression profiles from 1,037 high-throughput experiments for TCM herbs or ingredients. In addition, 12,933 targets and 28,212 diseases were further linked to 7,263 herbs and 49,258 ingredients by statistical inference via Fisher's exact test.

TCMIO

TCMIO [102] (Traditional Chinese Medicine on Immuno-Oncology, http://tcmio.xielab.net) is a comprehensive TCM-specific database for immuno-oncology (IO). Unlike the whole-scale herbs and ingredients in other databases, this database contains only formulae, herbs, ingredients, targets and diseases that are related to immuno-oncology.

1.3.2. Terminological system from Chinese terms

As TCM is a special system different from modern medicine, one of the greatest challenges in TCM study is how to correctly translate each TCM term into correspondent English so that international readers can understand the exact meaning. For instance, it is difficult to understand the term "quicken blood" in TCM. In fact, this term referred to the whole effect of "activating blood flow" or "promoting blood circulation". Thus, it is necessary to create a standardized system for TCM terminology [103].

Thanks to the developments of text mining technologies [104], researchers can standardize TCM terms and further build a bridge with modern medicine system, such as MeDisco/3T [105], ontology learning system [106] and TCMiner [107]. For example, a study has compiled 3,800 standard indication terms for translating TCM indication terms from Chinese to English [108]. Furthermore, the SymMap database (https://www.symmap.org/) was constructed to

provide a mapping relationship between manually extracted 1,717 TCM symptoms of TCM and symptoms of modern medicine [100]. This hints at the promising potential to integrate TCM with modern medicine at both the phenotypic and molecular levels. In addition, a clinical data warehouse (CDW) system was built by integrating clinical records from 20,000 TCM inpatient data and 20,000 outpatient data. These records included manifestations (e.g. symptoms, physical examinations, and laboratory test results), diagnosis and prescriptions, which would provide novel insight for TCM medical knowledge discovery and TCM clinical decision support. In addition, the CNA system called TCMNetBench [109] was also constructed as TCM clinical networks automatically, such as herb combination network, symptom co-occurrence network and complication disease network. In total, TCMNetBench covers 151 distinct modern diseases, 91 distinct TCM diseases and 216 distinct syndromes, suggesting that clinical TCM prescription databases are valuable for precision medicine study of TCM [105]. More importantly, an unified traditional Chinese medical language system (UTCMLS) [110] was also developed by an ontology approach, covering four basic top-level classes,14 sub-ontologies from the division of TCM discipline, 104 TCM semantic types as well as 59 kinds of semantic relationships between concepts and semantic types.

1.3.3. Target protein information

The pharmacology targets are closely related to the MOAs in TCM. Additionally, target protein is an important element for network pharmacology analysis [111]. However, the target information for the majority of TCM is unknown.

In the TCM, the validated ingredient-target interactions were mainly from four resources, including literatures, ChEMBL database [112], STITCH database [113] and HIT database [92]. For example, TCM-ID database and HERB database performed text mining from literatures. TCM-ID database, TCMAnalyzer database, and TCMIO database extracted bioactivity assay data from ChEMBL database while STITCH database. STITCH database provides a more comprehensive chemical-target interaction confidence score by considering all the information from literature mining, coexistence, concurrence, confusion. Compared with the prediction models for targets or docking methods [114], the interaction score is a more systematic characterization of chemical and target relationships. In addition, TCMSP collected compound-target interactions from the HIT database.

Except for validated targets, most of the targets for TCM ingredients are the putative targets from prediction. Furthermore, the prediction methods for ingredient-target association. For instance, TCM-ID and Database@taiwan TCM database utilized a docking method. TCM-ID applied ligand–protein inverse docking strategy- INVDOCK [114] method to search targets among human and mammalian proteins in the protein database Protein Data Bank (PDB) [115] to predict targets. INVDOCK is a ligand–protein inverse docking strategy by dock compounds to known ligand-binding pockets of each of the protein entries in the PDB database. Similarly, Database@taiwan also obtained targets by virtual screening by docking and molecular dynamics.

Nevertheless, those virtual screening by docking is time-consuming, limited to in target structure and relies on platform. Thus, target prediction models became popular in the research of TCM, such as the SysDT [116] model in TCMSP database, multi-voting chemical similarity ensemble [117] approach in YaTCM, MedChem Studio in ETCM database, balanced substructure-drug-target network-based inference (bSDTNBI) [118] approach in TCMIO, as well as Fisher's exact test [119] for statistical inference in HERB.

Overall, the target information in many TCM researches is mainly from prediction and the experimentally validated targets were not fully extracted from the literature or public target databases. In addition, the targets for herbs and formulae are usually considered as the absolute union of targets from related ingredients. However, the target for herbs is much more complex because of drug-drug interactions. Thus, specific target prediction models targeted at one herb or formula should be developed. For example, the similarity ensemble approach was proposed to associate herbs with their putative targets directly at the molecular level rather than merging all the targets of ingredients[120].

1.3.4. Herbogenomics

There have been various TCM databases only providing the information for formulae, herbs, ingredients, targets and diseases. However, TCM is a complex system that uses herb combination as a formula to treat diseases. It is a holistic system that relies on multiple targets and pathways. As a result, it is difficult to treat the targets for herbs and formulae as single compounds, which becomes a bottleneck for mechanism demonstration. With advances in

computer science, large-scale data generation, storage and analysis are also developing. In the medical field, omics data, such as chemmics, genomics and proteomics, are becoming a powerful resource for novel findings in medicine [121], which is called a data-driven strategy. Recently, an increasing number of research papers have performed high-throughput transcriptome experiments for ingredients, herbs or formulae to study the holistic molecular effects of TCM.

Genomics, as one of the most significant technologies, has been widely used in the biological fields [122]. Genomics also takes all gene sequence changes of the whole body as the research subjects. Those genes are interconnected with each other and work independently as complex system pathways, which is similar to holism in TCM. Therefore, genomics is suitable for exploring TCM effects at the molecular level. Herbogenomics [123] has been proposed as a novel concept—we can get the omics changes of the human body before and after treatment with TCM herbs, and further analysis could be done to explore the mechanism of the biological effects of herbs. Moreover, some novel approaches were also developed based on omics data to study TCM [124].

Normally, the transcriptional profiling of TCM includes the following steps: 1) Preparation of the extracts of herbs or formulae. TCM involves the oral administration of decoctions; 2) Apply extracts on objects at different conditions. Experiments objects can be cell lines, animals or patients; 3) RNA extraction and microarray processing; 4) Microarray data analysis; 5) Wet experiments validation [125]. For example, the DNA as a microarray-based gene expression value of the Chinese medicinal formula Si-Wu-Tang was used to reveal its phytoestrogenic activity on the cell line level [126]. Besides testing the gene expression data of herbs through experiments, we can also integrate them with existing big public databases. For example, the next-generation Connectivity Map resource was used to interrogate gene expression signatures and obtained 102 TCM components to prove the MOAs of TCM [127]. Recently, HERB database was constructed in 2020 to provide high-throughput experiments and references for herbs and ingredients in TCM. They extract the raw count from records about TCM and ingredients in Gene Expression Omnibus Database (GEO) [128] to perform quality control, filtering, normalization and alignment. In addition, differentially expressed genes (DEGs) were calculated between samples treated with ingredients or herbs and control samples to

characterise the perturbation in transcriptional level. After getting the upregulated and downregulated genes, the downstream pathway analysis and go function are enriched.

1.3.5. Metabolism

Metabolic profiling is the representation of path-physiological, pharmacodynamic and pharmacokinetic conditions after being influenced by environments, genetic factors or gut microbiome factors [129]. It was believed structural similarities between TOM-derived compounds and human metabolites can facilitate the identification of their MOAs and potentially reduce toxicities [50].

Chinmedomics is a multivariate data analysis method for herbal medicine using metabolite data [130]. This method could help researchers to understand the metabolites of integrated living systems and their dynamic responses to changes in both endogenous and exogenous factors. This approach has been applied successfully in TCM research [131, 132]. With a deeper understanding of TCM, it was found that some TCM herbs have pharmacological effects on the human body by adjusting the disordered gut microbiota environment under disease conditions. Thus, TCM-gut microbiota network pharmacology has become a "New Frontier" [133]. For example, many TCMs can be used as agents to prevent gut dysbiosis [134]. Dietary compounds and TCM have the bioactivity to ameliorate type-2 diabetes via regulating gut microbiota [135]. In addition, metabolites are the direct signature of phenotype symptoms, which can work as biomarkers for patients under different conditions [136]. It was also found that different tongue microbiota-imbalanced conditions are associated with Cold/Hot Syndromes in TCM [137].

1.3.6. Disease information

Target-based drug discovery (TDD) and phenotypic drug discovery (PDD) are the two main strategies for drug discovery [138]. There are many databases that provide ingredient–disease, herb–disease, and formula–disease data. However, the symptoms and disease classifications described in the TCM system are different from those terms in modern medicine, as they appear under the guidelines of TCM theories. As a result, most databases use a TDD method to understand TCM, based on the hypothesis that if a compound displays activity on disease-

related targets, it is a promising candidate to influence the disease. For example, the target–disease associations in TCM-ID database come from the Therapeutic Target Database (TTD).

Although TDD is popular for drug discovery, it is not perfect for TCM. This is mainly because most ingredient–target interactions in TCM databases are putative and thus the ingredient–disease association from TDD-based methods is undirected and predicted. In addition, the herb–disease and formulae–disease pairs should not be considered as simple unions of ingredient–disease but combinations with synergistic effects.

Considering the drawbacks of TDD approaches in TCM, the SymMap database was created to realise PDD drug discovery in TCM. SymMap mapped 1,717 TCM symptoms of 499 herbs in clinical practice to 961 modern symptoms as well as 5,235 modern diseases manually by highly experienced TCM practitioners. Furthermore, HERB manually extracted highly plausible diseases related to 394 herbs or ingredients from references.

1.4. Artificial intelligence (AI) application in medicine

Artificial intelligence (AI) [139], sometimes called machine intelligence, is intelligence demonstrated by machines to solve problems such as reasoning, knowledge representation, planning, learning, natural language processing, perception, and the ability to move and manipulate objects [140]. One of the branches of AI is machine learning, which is the scientific study of algorithms and statistical models that use computer systems to perform a specific task effectively. One advantage of machine learning is that it does not need explicit instructions but relies on patterns and inference instead. In addition, machine learning is powerful, especially when we face a complex problem involving a large amount of data and lots of variables. For example, massive amounts of imaging data (digital data adiology, pathology, dermatology, and ophthalmology) were generated in the medical area from patients or experiments [141].

1.4.1. Classifications of AI methods

Machine intelligence can also be classified into unsupervised learning [142] and supervised learning [143]. Unsupervised learning is to learn from only variables without considering the outcome (real value), while supervised learning is to learn from variables with real value as the

end point to refer to. For cluster methods, there are distance-based clusters such as K-means [144], K-Medoids [145], Fuzzy C means [146] and Hierarchical [147]. Probability-based methods are to obtain the probability distribution of input variables and the outcome result. Supervised learning [148] can also be classified into classification methods and regression methods based on whether the endpoint is the classification label or the continuous value. According to the underlying concepts, supervised learning can be classified into different types, including linear model, probability-based algorithm, instance-based algorithms, tree-based algorithms, and artificial neural networks.

Linear algorithms

Linear models are aimed at building a linear function between the variables with weight and the final endpoint. With this function, the endpoint of external samples can be predicted by fitting the variable feature into function. Traditional linear models include linear least-squares regression (LR) [149], logistic regression (LoR) [150], and linear discriminant analysis (LDA) [151].

Probability-based algorithms

Bayesian algorithms [152] are probability-based. For example, Naive Bayes' theorem assumes that the features are independent of each other and the final probability is the joint probability.

Instance-based algorithms

Instance-based methods classify the samples into groups based on the feature's similarity with instances in the training set. For example, as indicated in the name k-nearest neighbour (kNN) [153], kNN methods are assigned to a class by considering the closest k instances. The distance can be calculated by various distance measurements. Another popular instance-based method is the support vector machine (SVM) [154]. SVM is designed to achieve nonlinear separation in the original input space by projecting the data to a higher-dimensional space via kernel functions.

Tree-based algorithms

Tree-based methodology aims to build up a tree for classification with one node as one filtering rule based on the imported feature. Take decision tree (DT) [155], one of the most original treebased methods, as an example. The tree node is ordered by the importance of these features and the continuous dividing of samples into two or more subsets by this node (feature) until the samples in the subsets are pure enough (terminal node).

Random forest (RF) [156] is developed from DT. Compared with DT, RF is a parallel procession of individual DT by random sample selection as a training set. Finally, the final classification can be decided using majority voting on multiple trees (decision forest).

Artificial neural network algorithms

Artificial neural networks (ANN) resemble the working process of the neurons of the human brain: information from environments as the input layer; the brain starts to connect the neurons to transform the information, which is called hidden layers; finally, correct judgments for the complex information is returned (output layer) [157].

1.4.2. Application of AI methods in medicine fields

AI is becoming a major constituent of many applications in the medical field, including drug discovery, remote patient monitoring, medical diagnostics and imaging, risk management, wearables, virtual assistants and hospital management [158].

One of the good applications of supervised learning is that it is widely used in clinical medicine for prediction and diagnosis when there are already many record variables and corresponding labels from Clinical practice, especially the prediction of the patient's survival [159]. At present, a variety of AI models have been applied to help the diagnosis and prognosis for more accurate decision-making. Among various methods, the models of neural deep networks are difficult to interpret while models from the decision tree are most interpretable [160].

In addition, NNs have been widely applied to complex problems, especially those whose variables are huge. Recently, an inspiring study successfully predicted the protein structure using neural networks [161]. Protein is the fundamental element of humans to perform distinct

biological functions. As a result, structural prediction of proteins from their amino acid sequence can be used to determine the three-dimensional shape of a protein. Furthermore, deep convolutional neural networks are also state-of-the-art methods for medical image analysis [162]. For medical images, the realistic noise in them is usually mixed, complicated and sometimes unknown, which makes it a challenge to build up accurate models for image analysis.

1.4.3. Application of artificial intelligence in TCM

AI has been applied to solve complex problems in the TCM field [36, 163]. It was reported that there are 502 articles from 2000 to 2017 on machine learning models developed for TCM [164]. The application domain included symptom, syndrome differentiation, the efficacy of drugs and herbs. The most widely applied ML methods in TCM are BNs, ANNs as well as SVMs. There are various chemoinformatics, bioinformatics, and systems biology resources that can be used for drug–target networks construction of products [37]. It is increasingly popular to apply system biology approaches and machine learning techniques to complex biochemical and pharmacological datasets from TCM [34]. For example, a novel network-based methodology was developed to discriminate the potential effective drug combinations in clinical practice [35]. Clustering method was used to study the relationship between the hot or cold nature of herbs [38]. Unsupervised clustering method called a self-organizing map (SOM) was utilized to predict the cold and hot properties of herbs using fingerprints as chemical structure features. Moreover, the knowledge graph approach was promising to provide insight into the underlying mechanism by integrating comprehensive information in TCM together [165].

Tongue diagnosis [166] and pulse diagnosis [167] are important diagnostic methods in TCM, which are totally independent of the modern medicine system. Tongue diagnosis is to diagnose the disease or the condition of the body of a patient by looking at the texture of the tongue. Similarly, the pulse diagnosis is to diagnose patients by feeling their pulse beating at the measuring point of the radial artery. However, these diagnosis approaches require doctors to have rich experience to make the correct diagnosis [168]. Recently, an increasing number of image analyses have been performed to understand the tongue diagnosis or pulse diagnosis to make TCM diagnosis standardized. For instance, the approach of Bayesian networks was applied to explore the relationship between diseases and the chromatic and textural measures of tongue images from 455 patients affected by 13 common diseases, as well as 70 healthy

volunteers [169]. In addition, the tongue images of patients were used to predict clinical syndrome classification [170]. Also, a convolution neural networks model was developed to provide herbal prescriptions automatically based on features of tongue images [171].

1.5. System biological network and TCM

1.5.1. Cluster methods for TCM

Cluster methods are important analysis methods in network analysis to help find similar groups. In TCM, cluster methods are widely applied to classify patients into subclass diseases, which can be seen as one kind of precision medicine. For example, patients with the same disease named kidney deficiency by TCM system were diagnosed into different subclasses according to their syndrome using the cluster model of the latent tree model, suggesting that those resultant clusters can be used as the basis of syndrome differentiation [172]. Similarly, a subspace clustering algorithm was applied for TCM syndrome differentiation for another disease acquired immune deficiency syndrome (AIDS) based on their symptoms [173].

Except for syndrome differentiation, cluster methods can also be used to understand the special theories in TCM. For instance, herbs can be clustered into 5–10 categories based on the drug response, amplitude, and the occurrence of time. These categories illustrated a close relationship between tissue organs and TCM meridians. In addition, the heterogeneous information network method was proposed to cluster herbs with the symptoms, disease, formulae, and function as the main herb features [174]. Another advanced network clustering algorithm applied in TCM is the Random Walk algorithm, which starts from a random node (drug, target, or disease) and calculates the similarity between this node and its adjacent node to construct a “drug–target–disease” network heterogeneously [175].

1.5.2. System pharmacology applied on TCM

System pharmacology is a new concept that integrates system biology and pharmacology as a whole network. Thus, it has been widely applied in TCM for the exploration of active ingredients or targets and to understand therapeutic action mechanisms [176]. Generally speaking, the first step of network pharmacology analysis in TCM is to construct the

associations between five main entities, including formulae, herb, ingredient, target, and disease. Before 2014, multi-pharmacology networks contained herb–ingredients–target–disease associations. Gradually, the formulae–herb–ingredient–target–disease association network was further constructed to better understand drug combination in TCM formulae. Secondly, network analysis can be performed based on the associations to find common patterns or central important node nodes. Thirdly, kyoto encyclopedia of genes and genomes [177] (KEGG) biological pathways and Gene Ontology [178] (GO) functional terms can be enriched in various databases to discover potential MOAs of bioactivity effects. Finally, a complex formulae–herb–ingredient–target–disease-pathways network can be built up. The applied domain of network medicine in TCM include symptom differentiation, herb properties [19], herb combination [80, 179] and TCM diagnosis.

1.5.3. Network medicine for TCM symptom differentiation

Symptom differentiation is an important premise of precision medicine. At present, increasing efforts are being made to exploit TCM syndrome studies using network medicine methods, which facilitate TCM modernization and personalized medicine.

For instance, a complex network approach was proposed to map TCM symptoms to modern medicine symptoms [180]. In detail, 60 physical examination data on symptom terms as well as 199 TCM examination symptom terms were collected from 521 patients with fatty liver disease and 235 healthy subjects. The correlation between TCM and MM symptoms was calculated from the network. In addition, structural metrics are important to characterize network topology. Especially, the centrality metrics can be used to identify the significant symptoms. Finally, they found hub symptoms as biomarkers for fatty liver disease, including body mass index (BMI), low density lipoprotein, yellow tongue fur and thick tongue fur. This application illustrated the power of network medicine analysis

Pattern classification by network analysis is an important and effective way for symptom differentiation. For example, it was reported that patients with rheumatoid arthritis can be classified into two main TCM patterns—the cold pattern and the heat pattern according to TCM theory. The classification pattern was obtained by a series of network analysis based on the differentially expressed genes of patients with rheumatoid arthritis [181]. It was found that

people classified as cold patterns in TCM have similar biomarkers with each other and are different from patients among heat patterns. The same is for people in hot patterns. In addition, the network biomarkers of Cold and Hot Syndromes in TCM suggested that cold and hot networks might give rise to different clinical phenotypes.

1.6. The drug combination and TCM formulae

1.6.1. Drug combination

Many complicated diseases involve complex pathogenesis. Thus, it is difficult to use a single drug to treat such complex diseases with efficiency. Furthermore, some single drug treatments might cause some side effects as well as drug resistance [182]. In addition, recently, it was reported that the pharmaceutical industry started to get fewer drugs than before by more investment in the discovery of novel drugs. Considering all the above drawbacks of single drug strategies, drug combinations have been becoming increasingly popular because of their synergistic effects and fewer side effects.

At present, many high-throughput screening methods have been applied to identify potential drug combinations and achieve some success. However, searching for the potential synergistic combinations among numerous compounds is impossible because wet experiments are time-consuming, labour-intensive and expensive [11]. In contrast, the application of computational tools to immense available pharmacological data offers an alternative way to identify potential drug combinations. In contrast, the application of computational tools to immense available pharmacological data offers an alternative way to identify potential drug combinations. Especially, omics data could be used to represent the perturbations that a drug might raise for biological pathways or molecular interactions and thus help to understand the process systematically, which is called network pharmacology [4-6]. Multi–target drugs or drug combinations are designed to influence multiple disease-causing genes rather than focus on only one individual gene [7-10]. To realize network pharmacology paradigm, particular computational models should be developed to measure the perturbation of drug combinations on robust disease phenotypes by targeting multiple biological pathways [35, 54, 183]. Furthermore, compared to individual drug treatment, side effects in drug combination can

decrease accordingly by the reduced dose of individual constituents but achieve the same or stronger therapeutic effects [15].

Among various systematic methods, network pharmacology approaches have attracted considerable attention because of their potential for the understanding of drug interactions in many complex diseases more recently. For instance, the DrugComboRanker approach was developed to prioritize the most potential synergistic drug combinations by their functional protein–protein network [184]. Moreover, network-based methodology can be used to characterise the distance between two drugs in terms of target distributions among the protein–protein network (PPI) [35]. They elucidate that drug combinations approved clinically are closer in PPI network than random drug pairs. Moreover, a modularity analysis and multipartite networks model might be a suitable approach for exploiting the MOAs of traditional medicine [185]. For example, the system pharmacology approach has been used to study the MOAs for blood stasis syndrome in the herb pair *Angelica sinensis–Flos Carthami* [186].

1.6.2. TCM formulae and herb pairs

As the empirical paradigm of the drug combination, TCM formulae have developed for thousands of years and exhibit obvious synergistic effects. Hence, a deep understanding of TCM formulae might provide new insight for drug combination [13]. In the long history of the development of TCM, it was found that using several herbs together to relieve complex symptoms would achieve better effects than using single herbs. In modern medicine, concepts of system biology and drug combination were introduced because single drugs show the disadvantages of drug resistance and side effects. Similarly, TCM formulae can prevent and treat diseases at system biological levels, and herbs in formulae possess synergistic effects [14]. Thus, TCM is a valuable resource for drug combination discovery. In fact, many famous formulae have been approved as drugs on the market. One good example is the Fufang Danshen Diwan (Dantonic pill), which is approved for the market in many countries for cardiovascular diseases [16]. Dantonic pill is a botanical drug consisting of only *Radix Salviae Miltiorrhizae* (Danshen) and *Radix Notoginseng* (Sanqi).

Herb pair is an important concept in TCM as TCM formulae were gradually developed from two herbs as the fundamental combinations to multiple herbs. A patient with the same diseases

might have different symptoms [4, 40, 41], which need more precise TCM prescriptions based on the condition of patients. For instance, the common herb pair *Coptis chinensis* (Huang Lian) and *Evodia rutaecarpa* (Wu Zhu Yu) has been used along with other herbs for many disease indications [187], including the inhibition of *Heliobacter pylori* [188] and the treatment of obesity [189]. The understanding of the underlying mechanism of herb pairs would hold the key to the discovery of novel drug combination as herb pairs are fundamental elements of the TCM formula [33, 44]. Recently, many computational methods have been developed to investigate the herb pairs and herb formulae. For example, Yu *et al*. found that TCM-herb properties can be used to discriminate known TCM herb pairs from non-TCM herb pairs by computational methods, including probabilistic neural network (PNN), k-nearest neighbour (kNN), and support vector machine (SVM) [41]. Another study revealed the chemical and pharmacological characteristics of herb pairs by the herb feature mapping in a network pharmacology model [40]. Notably, genetic algorithms [42] were applied for the discovery of herb combinations to treat lung cancer [43]. They showed that the predicted herb formulae tend to have higher degrees of connections in one herb–herb frequency network, suggesting the feasibility of using machine learning and network modelling for understanding herb–herb interactions.

Despite these initial efforts, it is still difficult to elucidate the underlying MOAs of the stronger therapeutic effects of herb pairs as well as their synergistic effects as TCM formula is a complex system and normally involve multiple herbs with multiple ingredients [48, 49, 183]. Recent studies have shown that synergistic effects of herb combinations might stem from the interactions between their ingredients [50]. Additionally, ingredients within a herb might also interact synergistically to play plo-pharmacological effects. To take a case in point, the synergistic antioxidant activity of Ginseng (the root of *Panax ginseng*) is caused by ginsenoside Rb1, ginsenoside Rg1 and ginsenoside 20(S)-protopanaxatriol [53]. The studies above focus on specific herbs and can be seen as the basis for a systemic model to characterize the interactions of TCM herbs at the molecular level.

1.7. Challenges of TCM

Although TCM theory is self-consistent as a whole system and is independent of the modern medicine system, the MOAs of the theories and therapeutic effects remain unclear. To illustrate, the five elements and qi refer to more metaphysical concepts rather than physical evidence, which makes it difficult to be interpreted by modern physiological or medical entities [18]. Furthermore, biological targets for the majority of TCM ingredients are missing. How to define or characterise the target of one herb at a holistic level is still a big challenge in front of the modernization of TCM. More importantly, the toxicity caused by TCM and unknown concentration of TCM are also problems that impede the development of TCM.

1.7.1. Toxicity caused by TCM

Safety is always one of the most important principles for medicine. With the widespread application of TCM in many countries for disease treatments, the toxicity of TCM has gradually been reported and increased intensive concern [190]. Some toxicity caused by TCM can be seen as adverse effects, like many drugs on markets. However, as TCM is a complex system different from modern systems, the regulations and clinical research are limited compared with modern systems, which might lead to severe toxicity. At present, the toxicity of TCM is a critical problem and hinders the development and globalization of TCM. Thus, it is necessary to figure out the potentially toxic ingredients and understand the MOAs of toxicity. For example, advanced quality detection methods should be introduced to study the structure and bioactivity of ingredients in TCM medicine, such as TLC (Thin-Layer Chromatography) [191] and HPLC-fingerprint technique (High-Performance Liquid Chromatography) [192]. Also, the Good Agricultural Practices (cGAP) and Good Manufacturing Practices (cGMP) should be practiced during the growth and preparation period of TCM to avoid any external toxicity from pollution from the environment. Thirdly, the clinical biomarkers should be detected to help timely TCM toxicity diagnosis and treatment [193].

1.7.2. Ingredients in herbs and concentrations of ingredients

In the future, to achieve optimal precision medicine using the formulae of TCM, we need to understand the TCM in quantification ways. Quantification includes two aspects: 1) the

concentration of ingredients in herbs; 2) the percentage of herbs in one formula. However, it is difficult to get the exact percentage of herbs in one formula because the percentage of herbs in one formula may vary according to the different patient's syndrome by doctor diagnosis [194]. In addition, it is challenging to determine the concentration of ingredients or the biological activity ingredients because ingredients can be influenced by the gut microbiota [195]. Indeed, increasingly researches have shown that the original ingredients of herbs may not be absorbed directly but further be processed by gut microbiota in the intestines. In this process, herbal ingredients interact with gut microbiota and are transformed into active metabolites. Thus, it hints us at the importance of microbiota analyses by modern computational approaches.

2. AIMS OF THE STUDY

This thesis aimed to build up a bridge between traditional medicine and modern medicine using multiple bioinformatics tools and further to better understand the underlying mechanism. With a better demonstration of the mechanism, we believe that the application of traditional medicine can be extended widely to many fields of disease, especially complex ones, and complement the disadvantages of modern medicine. Considering that traditional medicine is a different system compared with modern medicine, a molecular-level explanation for its unique and fundamental theories and principles is also vital for TCM modernization. In addition, some valuable principles in TCM, "Take the human body as a holistic system and use TCM formulae to treat disease", are consistent with the emerging popular theory–drug combinations and multiple pharmacology after people gradually realise the challenges of modern medicine. To bridge TCM with modern medicine at the ingredient level, integrated information for TCM from TCM databases and other biological public databases should be extracted.

Thus, our study consisted of three parts with specific objectives:

1) To investigate the herb–meridian classification of TCM by machine learning methods using the structures of the ingredients and their ADME properties.

2) To build network-based models to explore the mechanism of herb combination in TCM and further apply this model to discover novel drug combinations from TCM.

3) To explore the associations between herbs or ingredients and their important biological characteristics, such as property, meridian, structure, or target via clusters from community analysis of the multipartite network.

3. MATERIALS AND METHODS

3.1. Data collection (I, II, and III)

3.1.1. Herb–ingredient relationship and ingredient structures (I, II, and III)

Ingredients in TCM are not only important constituents of natural products but also valuable resources for drug discovery. As we know, the majority of TCM theories are mainly constructed at the herb level because of a lack of knowledge at the molecular level in ancient times. Thanks to the development of modern technology, the ingredients in herbs, as well as their target interaction with the human body were uncovered. To understand TCM at a molecular level, ingredients can be seen as the bridges to modern medicine systems. Particularly, the target and structure information of ingredients can be further used for pharmacology studies.

We extracted the ingredient information of herbs from the TCM database called TCMID. TCMID is one of the largest databases of TCM, providing 49,000 prescriptions, including 8,159 herbs and 25,210 ingredients. Only those ingredients with clear structure information were kept for further study.

3.1.2. Herb–meridian association of herbs (I)

Meridian's classification was concluded from thousands of clinical practices by the therapeutic effects of herbs for various diseases, which might be related to different parts of the human body. We collected the herb–meridian associations from the TCMID database and further confirmed these associations by the TCM ancient books. Only the meridians consistent between the records from TCMID and books were kept. According to TCM meridian theories, there are 12 principal meridians. However, considering that the aim of our study **I** is to build reliable machine learning models for meridian prediction, those meridians distributed by only fewer herbs will not be included when training the meridian models.

3.1.3. ADME properties of ingredients (I)

ADME properties refer to all the processes the compounds might undergo in the human body and mainly concern four aspects: absorption, distribution, metabolism and excretion. ADME properties can influence the pharmacokinetics of a compound. Therefore, to avoid side effects and improve efficiency in drug discovery, ADME is a key factor to consider as the optimization processes. As a result, we also took the ADME properties of the ingredients into consideration as features in study **I**. The SwissADME database [196] was utilized to calculate the 36 ADME properties (six drug-likeness-related from different models, four medicinal chemistry, nine pharmacokinetics, nine physicochemical-related, three water solubility properties–related as well as five lipophilicity-related).

3.1.4. TCM formulae–herb relationship and herb pairs (II)

We managed to obtain 46,929 herbal formulae, and the related herbs consist of the formulae from TCMID, including 16,767 herbs. As herb pairs are the most fundamental component in the formulae, we calculate the herb pair frequency by the times of their coexistence in one formula, and the pairwise 349,197 herb pairs were obtained for further study. As one herb pair may be used to play distinct roles in the different herbal formulae, herb pairs with frequencies larger than 358 were selected for further study. On the other hand, we randomly generated herb pairs that did not exist in any actual herbal formulae as random herb pairs for comparison as top herb pairs. To validate the network model with an independent set, 268 herb pairs were extracted from traditional medicine literature as external dataset.

3.1.5. Targets of ingredients (II and III)

The target information of ingredients indicate the potential pharmacological bioactivities of ingredients. Unfortunately, the target information for most ingredients is missing due to the expensive cost of target identification. To overcome this problem, we extracted the ingredient–target interactions from the STITCH database [113] with the chemical–target interactions from both experimental evidence and text mining. Herbs and ingredients without target information were excluded.

3.1.6. Protein–protein interactions (II)

Finally, 243,603 protein–protein interactions among 16,677 proteins were collected from 11 resources, including MINT [197], IntAct [198], InnateDB [199], PINA [200], HPRD [201], HI-II-14_Net [202], PhosphositePlus [203], KinomeNetworkX [204], Instruct [205], SignaLink [206] and BioGRID [207].

3.2. Fingerprint calculation for ingredients (I)

To investigate the meridian classification using machine learning methods for TCM herbs, we determined that the fingerprint was used as a prominent feature to characterise the structural information of herbal ingredients. The canonical SMILES [208] is standardized by Open Babel [209] and further transformed into four kinds of fingerprint features using PaDEL-Descriptor [210] software separately, including Ext [211], PubChem [212], MACCS [213], and Sub [214]. Fingerprint is a string of binary bits to represent the particular substructure characteristics in a reference library. For instance, the PubChem fingerprint from the PubChem database contains 881 bits, and the MACCS fingerprint from MDL Company consists of 166 bits. There are 307 and 1024 bits in the Sub fingerprint and Ext fingerprint respectively.

3.3. Machine learning models (I)

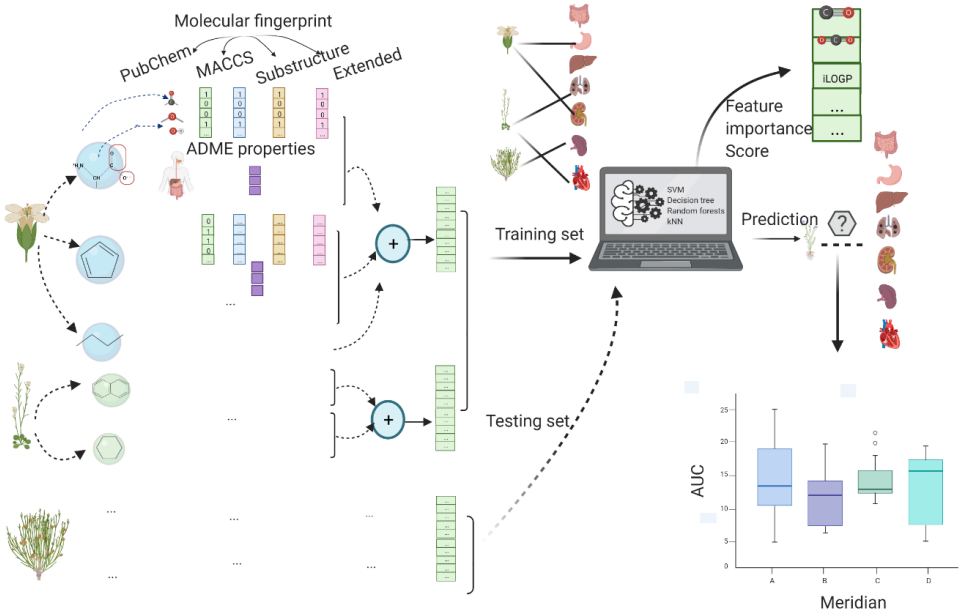

Figure 1. Workflow of the study **I**. Firstly, the meridian and ingredients information of herbs were collected. Secondly, the four types of fingerprints and ADME properties were calculated as features for ML. The herb feature is derived from the molecular feature by the average feature of all the contained ingredients. The meridian of the compounds is from the union meridians of all the related herbs. Thirdly, the whole dataset was split into training sets for model training and testing sets for model evaluation. The machine learning models were trained by parameter optimization, model selection, and feature selection. Finally, the important score was calculated to find the most important features relative to meridians.

3.3.1. Construction of compound-feature matrix and herb-feature matrix

The features of a compound were integrated using various fingerprints and ADME features. In detail, there are 2,378 fingerprint features, including 1024 Ext, 881 PubChem, 307 Sub, and 166 MACCS and 36 ADME property features. The machine learning models were first trained on different datasets using four fingerprint types separately to compare their performance. After determining the best fingerprint type, ADME features were added into the best fingerprint type to check the change of model performance. Finally, X_C matrix was prepared as the input of machine learning (ML) with 2,414 features for each of 10,053 compounds.

In the study of drug combination, the feature of two drugs is usually considered the sum of the features of both drugs. In the same way, a herb is a mixture of multiple ingredients. As a result, herb features can be obtained as below:

Let $C_j = (c_1, c_2, \dots, c_k)$ denote the set of ingredients in herb j, where k means the number of compounds. The compound feature is defined as $F_{compound} = (f_1, f_2, \dots, f_n)$, in which n is the total number of features. Then, the herb feature $F_{herb} = (g_1, g_2, \dots, g_n)$ is the average feature value of all compounds, i.e.

$$g_{i(i=1,\dots,n)} = \frac{\sum c_1, c_2, \dots, c_k f_i}{k} \quad (1)$$

To be specific, for all 646 herbs collected, the feature can be determined by the 2,414 features of their ingredients. The 646x2414 herb-feature matrix (HF) is defined as:

$$\mathbf{HF} = \begin{matrix} \\ F_1 \\ F_2 \\ F_3 \\ F_4 \\ \dots \end{matrix} \begin{bmatrix} 0.2 & 0.1 & 0.3 & 0 & 0 \\ 0 & 0.1 & 0.1 & 0 & 0.8 \\ 0.1 & 0.6 & 0 & 0.1 & 1 \\ 0.5 & 0 & 0.1 & 0.3 & 0.1 \\ 0 & 0.4 & 0.2 & 0 & 0 \end{bmatrix}_{646\times2414} \quad (2)$$

As ADME properties have a close relationship with the biological effects of ingredients, to evaluate the influence of ADME, we also test the datasets with samples whose feature is only derived from the compounds with good ADME properties, mainly by logS ≤ -6, in all the three water solubility models and gastrointestinal absorption ≤ 30%. Accordingly, after the ADME

filtering, 583 herbs and 4,922 compounds were kept to compare the model performance with the models before filtering.

3.3.2. Construction of herb–meridian matrix and compound-meridian matrix

The meridian vector of herbs is denoted as $M_{herb} = (m_1, m_2, \dots, m_{12})$. Thus, the herb–meridian matrix (HM) of the 646 herbs among the 12 meridians is as below:

$$\mathbf{HM} = \begin{array}{c} \\ \mathrm{M_1} \\ \mathrm{M_2} \\ \mathrm{M_3} \\ \mathrm{M_4} \\ \dots \end{array} \begin{array}{c} \begin{array}{ccccc} \text{Lung} & \text{Spleen} & \text{Stomach} & \text{Kidney} & \dots \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 \end{bmatrix}_{646\times12} \end{array} \quad (3)$$

$H_j = (h_1, h_2, \dots, h_p)$ is a set of p herbs with this compound. Similarly, the meridian vector of the compound is $M_{compound} = (l_1, l_2, \dots, l_{12})$ and can be calculated by merging the meridians of the herbs in H_j, i.e.

$$l_{i,i=1,\dots,12} = I(\textstyle\sum_{h_1,h_2,\dots,h_p} \quad m_i > 0) \quad (4)$$

$I(\cdot)$ is an indicator function. The full compound-meridian (CM) matrix (compounds = 10,053 and meridians = 12) was prepared as below:

$$\mathbf{CM} = \begin{array}{c} \\ \mathrm{C_1} \\ \mathrm{C_2} \\ \mathrm{C_3} \\ \mathrm{C_4} \\ \dots \end{array} \begin{array}{c} \begin{array}{ccccc} \text{Lung} & \text{Spleen} & \text{Stomach} & \text{Kidney} & \dots \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}_{10053\times12} \end{array} \quad (5)$$

3.3.3. Training the machine learning models

In the machine learning models, the endpoint of each meridian is binary, where "1" means a herb was recorded to be classified into a specific meridian while "0" means it was not. Four supervised ML methods SVM, DT, RF and kNN [215] were applied for the prediction of meridians by the R package caret [216]. SVM searches out a hyperplane to separate samples from different classes into individual space from a hyperplane, which achieves minimal error. DT resembles a decision tree with observations to determine the branch node. kNN methods

can classify the samples by a majority vote of its closest k neighbours, according to the distance by features. RF is based on tree votes by multiple decision trees, and the majority voted one as the final classification. To make the model robust, five-fold cross-validation [217] was utilized to avoid overfitting by samples split randomly into 70% as training sets and 30% as testing sets. The training samples were further split into five folds with the same size. Four folds were used for training, and the reminding one was kept for evaluation iteratively for each fold; thus, the process was repeated five times. After deciding the best model with the highest accuracy, the remaining 30% samples can be fit into models for external evaluation. To benchmark the model performance for each meridian, a permutation was adopted by randomly shuffling meridian labels according to the ratio of positive and negative samples, which can be seen as the baseline. The R scripts and input data are available at https://github.com/herb-medicne/meridian-prediction.

3.3.4. Evaluating the prediction accuracy for meridians

We applied a confusion matrix to do model evaluation. The overall prediction accuracy is defined as:

$$overall\ accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (6)$$

True positive (TP) indicates the number of positive samples (i.e. herbs or compounds) that are correctly identified for a given meridian, while false positive (FP) is the number of negative samples that are wrongly identified as positive. It was the same for true negative (TN) and false negative (FN). Balanced accuracy is defined as an evaluation metric of the average sensitivity and specificity to adjust the inflated overall accuracy for imbalanced data:

$$Balanced\ accuracy = \frac{\frac{TP}{TP+FN}+\frac{TN}{FP+TN}}{2} \quad (7)$$

Besides the evaluation mentioned above, the area under the receiver operating characteristic curve (AUROC) and Matthews correlation coefficient (MCC) [218] were also calculated. MCC is defined as:

$$MCC = \frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (8)$$

AUROC is defined as:

$$\text{AUROC} = \int_{x=0}^{1} \text{TPR}\left(\text{FPR}^{-1}(x)\right)dx \quad (9)$$

In AUROC, the true positive rate (TPR) is $\text{TPR}(t) = \int_t^{\infty} f_1(x)dx$ while the false-positive rate (FDR) is $\text{FPR}(t) = \int_t^{\infty} f_0(x)dx$ for a given classification threshold t. $f_1(x)$ refers to the probability density functions for the predicted score for an instance if it is positive, and $f_0(x)$ is the negative class.

3.3.5. Identification of key features for the prediction of meridians

To measure the contribution of each feature to the accurate prediction of the meridian classification, the varImp package [219] was adopted to score the features by their importance. Additionally, we used SARpy [220] to make statistics on the frequency of substructures and to find the key features for meridian prediction. The significance of each substructure is evaluated as the likelihood ratio by SARpy:

$$likelihood\ ratio = \frac{TP/FP}{P/N} \quad (10)$$

Among all the compounds classified into one meridian, compounds with a specific substructure is TP. The left compound that does not belong is FP. The top 10 substructures by the likelihood ratio score were selected as key features to analyse their association with meridians.

3.4. Network proximity model definition of herb pairs (II)

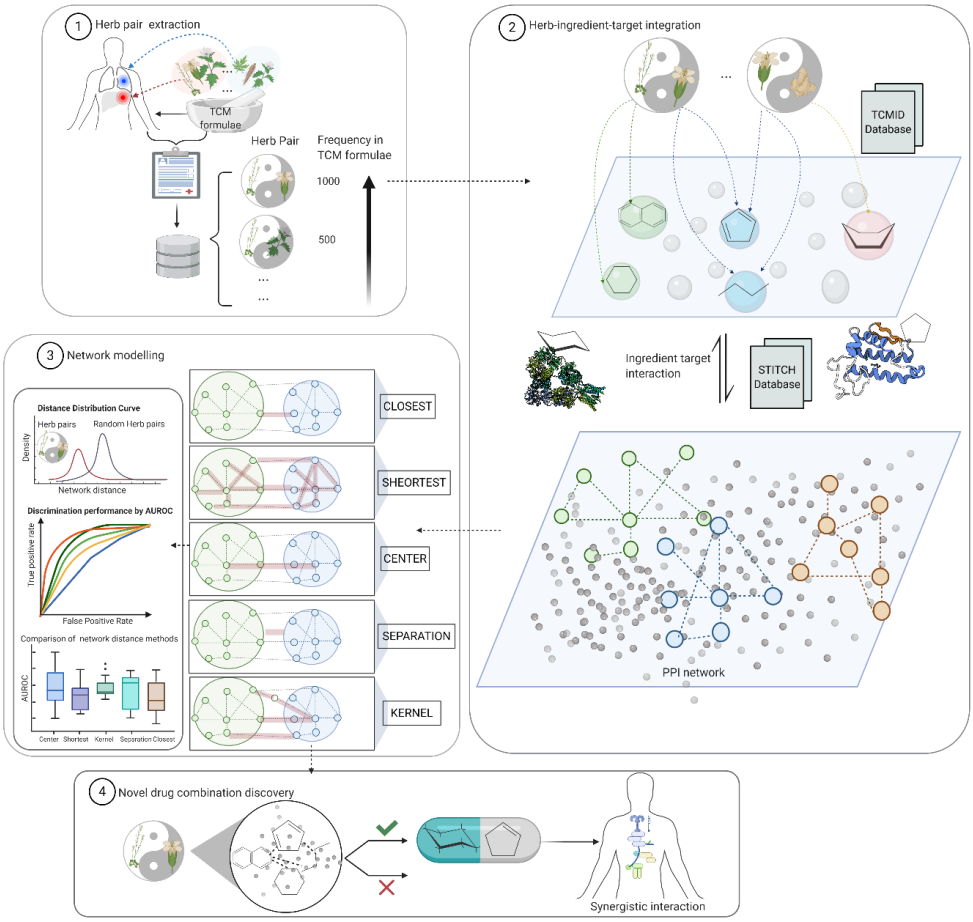

Figure 2: Workflow of the network construction between herb pairs based on interaction. Firstly, top frequent herb pairs were derived from existing herbal formulae by statistical analysis. Secondly, herb–ingredient, ingredient–target and target–target interactions were extracted from a public database. Thirdly, the network proximity of two ingredients can be calculated among PPI networks at the target level, and further distance between two herbs was calculated by the ingredient–ingredient distance at the ingredient level. In total, there

were five types of common network proximity methods to make comparisons to find the best model, including closest, shortest, separate, kernel and center distances. Finally, the best distance metrics that can discriminate herb pairs will be applied for novel combinational drug discovery.

Among various systematic methods, network pharmacology approaches have gained increasing interest for their wide application in the field of drug interactions in various complex diseases. For example, DrugComboRanker was proposed to prioritize potential novel synergistic drug combinations with the PPI network as the background network to do calculations [184]. Additionally, the associated network and targets can be used to explore their MOAs. Cheng *et al*. proposed PPI-based approaches to characterise a good drug combination [35]. Like many successful application cases in exploring the relationship between chemicals and diseases, network-based models could also be potential for rational TCM herb interactions.

In this study, the hypothesis is that using network pharmacology models to understand the drug–target interactions might provide novel insights into the underlying MOAs of herb combinations [132, 221, 222]. The frequencies of herb pairs in TCM herb formulae were calculated. Network-based models were built up to characterise the distance of herbs within a herb pair in a PPI network. The model applied five distance metrics (the closest, shortest, separate, kernel and center) separately to define the interactions of herbs at the herb, ingredient, and target levels. In addition, AUPRC and AUROC were used to evaluate models and thus help to decide the best distance methods that can discriminate the top frequent herb pairs against random herb pairs.

3.4.1. Construction of network proximity models for herb pairs

The herb–herb distance was calculated with ingredients as the nodes in which the distance between a pair of ingredients can be further obtained with their target as nodes in the PPI network. Given that $I(A) = (a_1, a_2, \dots)$ is a set of ingredients of herb A, while $T(a) = (t_1, t_2, \dots)$ is a set of targets for ingredient a, the same is for herb B. Five measures from Cheng *et al*. were utilized to determine the distance between two herbs, including closest, separation, shortest, kernel and center [35].

The closest distance is defined:

$$d_{I(A)I(B)}^{closest} = \frac{1}{||I(A)|| + ||I(B)||}\left(\sum_{a\in I(A)} min_{b\in I(B)} di(a,b) + \sum_{b\in I(B)} min_{a\in I(A)} di(a,b)\right) \quad (11)$$

$di(a,b)$ is the distance between two ingredient nodes in herbs A and B. ||I(A)|| and ||I(B)|| represent the numbers of ingredients for herbs A and B, respectively. Taking ingredients in herb A as starting points, all the pairwise distances of the ingredient in herb B were calculated, and the minimal distance was selected as the closest distance for the starting ingredient. Finally, the mean value of the closest distance for all the ingredients in herbs A and B was considered as their closest distance $d_{I(A)I(B)}^{closest}$.

The separation distance is developed from the closest distance using the average closest distances between A and B separately subtracted from the closest distance between A and B:

$$d_{I(A)I(B)}^{separation} = d_{I(A)I(B)}^{closest} - \frac{d_{I(A)I(A)}^{closest} + d_{I(B)I(B)}^{closest}}{2} \quad (12)$$

To obtain the shortest distance, the sum value of pairwise ingredient distances between nodes in herbs A and B was obtained, and this value was further normalized by the product of their sizes:

$$d_{I(A)I(B)}^{shortest} = \frac{1}{||I(A)|| \times ||I(B)||}\sum_{a\in I(A), b\in I(B)} di(a,b) \quad (13)$$

The kernel distance is the average exponent-based pairwise distance, which is normalized by their relative network sizes:

$$d_{I(A)I(B)}^{kernel} = \frac{-1}{||I(A)|| + ||I(B)||}\left(\sum_{a\in I(A)} ln \sum_{b\in I(B)} \frac{e^{-(di(a,b)+1)}}{||I(B)||} + \sum_{b\in I(B)} ln \sum_{a\in I(A)} \frac{e^{-(di(a,b)+1)}}{||I(A)||}\right) (14)$$

In the center distance methods, we first determined the nodes with a minimal sum of distances as the center nodes of A and B, separately. Then, the distance between the two centers was considered the final center distance:

$$d_{I(A)I(B)}^{center} = di\left(center_{I(A)}, center_{I(B)}\right) (15)$$

$$center_{I(A\ or\ B)} = argmin_{u \in I(A\ or\ B)} \sum_{b \in I(B\ or\ A)} di(b,u)\ (16)$$

The equations above were defined to calculate the distances between two ingredients $di(a,b)$. Indeed, five kinds of network distance can be referred to by their target profiles T(a) and T(b), including:

$$di_{(a,b)}^{closest} = \frac{1}{||T(a)|| + ||T(b)||}\left(\sum_{i \in T(a)} min_{j \in T(b)} dt(i,j) + \sum_{j \in T(b)} min_{i \in T(a)} dt(i,j)\right) (17)$$

$$di_{(a,b)}^{separation} = di_{T(a)T(b)}^{closest} - \frac{di_{T(a)T(a)}^{closest} + di_{T(b)T(b)}^{closest}}{2} (18)$$

$$di_{(a,b)}^{shortest} = \frac{1}{||T(a)|| \times ||T(b)||} \sum_{i \in T(a), j \in T(b)} dt(i,j)\ (19)$$

$$di_{(a,b)}^{kernel} = \frac{-1}{||T(a)|| + ||T(b)||}\left(\sum_{i \in T(a)} ln \sum_{j \in T(b)} \frac{e^{-(dt(i,j)+1)}}{||T(b)||} + \sum_{j \in T(b)} ln \sum_{i \in T(a)} \frac{e^{-(dt(i,j)+1)}}{||T(a)||}\right) (20)$$

$$di_{(a,b)}^{center} = dt\left(centre_{T(a)}, centre_{T(b)}\right) (21)$$

Therefore, five distance methods were performed at both the target and the ingredient levels. In total, 25 different herb distances were utilized with an exhaustive combination of five network methods. Let us take the model with the closest methods at the ingredient level and the closest methods at the target level as an example. The closest distance for two herbs at the ingredient level can be defined as:

$$d_{I(A)I(B)}^{closest} = \frac{1}{||I(A)|| + ||I(B)||}\left(\sum_{a \in I(A)} min_{b \in I(B)} di(a,b) + \sum_{b \in I(B)} min_{a \in I(A)} di(a,b)\right) (21)$$

$di(a,b)$ for ingredient a and ingredient b is defined as:

$$di_{T(a)T(b)}^{closest} = \frac{1}{||T(a)|| + ||T(b)||}\left(\sum_{i \in T(a)} min_{j \in T(b)} dt(i,j) + \sum_{j \in T(b)} min_{i \in T(a)} dt(i,j)\right) \quad (22)$$

$dt(i,j)$ is the shortest path length between the two targets in the PPI network.

3.4.2. Evaluating the discrimination performance of the proximity distances

To evaluate the discrimination of network models for herb pairs with synergistic effects, AUROC was utilized with 200 top herb pairs as positive samples and 200 random ones as negative samples. There is other assessing criteria, such as true positive rate or recall rate. However, AUC is an overall evaluation that consider different thresholds. From the AUC for ROC curve, we can see the best threshold for classifying positive and negative with the most true positive and the least false positive. In addition, AUC is robust for the imbalance samples. To make the model robust, the process was repeated around 50 times to obtain the average AUROC and AUPRC. In addition, 268 herb pairs were extracted from the literature as external datasets to further validate the network models.

3.4.3. Case study of the herb pair Astragalus membranaceus and Glycyrrhiza uralensis

Huangqi decoction is a TCM formula consisting of two herbs: *Radix Astragali* (Huang Qi) and *Radix Glycyrrhizae* (Gan Cao). It is reported that the Huangqi decoction can be used to treat liver fibrosis [223] and cirrhosis disease [224], while neither *Radix Astragali* nor *Radix Glycyrrhizae* can achieve same effective alone. Therefore, it is important to identify the synergistic effects between them and explain why they may be used to treat liver diseases as one herb pair.

Based on the best network model we have decided, the distances between *Radix Astragali* and *Glycyrrhizae* were calculated to compare to the distance of a random herb pair. Furthermore, we also calculated the center ingredients by the center methods. To exploit the potential MOAs for liver fibrosis of center ingredients, the minimum local PPI network from STITCH was set up by all the targets of the center ingredients. Pathway enrichment analysis was performed by enrichr [225] with the proteins in the local PPI network.

3.5. Multipartite network models for understanding Traditional Medicine (III）

3.5.1. Construction of bipartite network

In study **III,** using a modularity analysis[55] of multipartite networks[58], we aimed to further illustrate that the classifications in TCM theories have close association with the chemical properties of herb ingredients. Herbal medicines and chemical ingredients were extracted from the TCMID database. The herbal medicines and ingredients are the two parts of a bipartite network. Afterwards, the disconnected nodes were removed to build a component of the graph.

3.5.2. Construction of multipartite network and community detection

Next, we verified our hypothesis that the herbs or active ingredients that are clustered into the same community are more similar in their features. In other words, the ingredient similarity Network (ISN) was constructed with ingredients as nodes and edges if two ingredients share at least one more natural product. Similarly, we reconstructed the natural product (TCM herbs) similarity network (NSN) by linking natural products if they have at least one common ingredient. Then, the communities were detected from ISN and NSN respectively by optimizing a modularity score [226].

3.5.3. Similarity analysis of four types of features among communities

Moreover, our study was based on four types of features, including meridians and properties of the natural products as well as protein targets and SMILE string of ingredients. Meridians and properties were collected from TCMID. Furthermore, The SMILE strings of these ingredients were from PubChem and identified or predicted protein targets were extracted from STITCH databases. The average pairwise intersection of meridian and property profiles was computed separately for each cluster in NSN. Likewise, the similarity of ingredients in each cluster was defined as the average Dice index of the SMILE string as well as the pairwise intersection of their protein targets.

4. RESULTS AND DISCUSSION

4.1. Predicting meridians by machine learning approaches

4.1.1. Distribution of meridians at the herb level and the compound level

In total, 49,000 formulae, 8,159 herbs and 25,210 ingredients were collected from the TCMID database. Herbs without meridian information were excluded from further study and 464 herbs were left.

To investigate the constitution of the meridian among herbs, we performed a data analysis for the distribution of meridians at both herb and ingredient levels. At the herb level, 333 herbs were classified as Liver meridian, followed by Lung (n=237), Heart (n=155), Stomach (n = 235), Spleen (n=213), Kidney (n=181), and Large Intestine (n=111) (**I**, Fig 2A). However, the other meridians are limited by the number of herbs, as there are less than 60 herbs in each of the five meridians: Bladder, Gallbladder, Small Intestine, Cardiovascular and Three End. Considering that unbalanced datasets might lead to over-interpretation in machine learning models, we focused only on the former meridians.

We further analysed common herbs between different meridians. In fact, 89.8% of 646 herbs were classified into more than one meridian although the overlap rates that might vary between meridians. Kidney and Liver shared the largest number of herbs, with 51 in common. Furthermore, there are 36 overlapped herbs between Liver and Heart, followed by 30 herbs between meridian Liver and meridian Stomach (**I**, Fig 2A). In addition, we found that all the herbs from Stomach, Spleen or Large Intestine meridians have the tendency to have multiple meridian classification according to TCM records. On the other hand, there are fewer herbs that belong to only one meridian. Similar patterns were also observed at the compound level (**I**, Fig 2B). The high meridian overlapped rate among herbs demonstrated its multi–target characteristics through influencing different tissues of the human body.

Besides the number of common herbs, we also utilized Jaccard coefficients [227] to evaluate the similarity between different meridians. The Spleen and the Stomach are most similar to each other, with Jaccard indexes of 0.31 and 0.42 at the herb level and the compound level

separately. In contrast, the Heart and the Large Intestine are quite different from each other among the meridian distribution, with 0.04 and 0.14 at the herb and compound levels, respectively (**I**, Fig 2C-D). The average pairwise Jaccard index of the herb–meridians is 0.15. Likewise, the average Jaccard index is 0.26 for meridians of compounds. In short, a lower Jaccard value indicates weaker correlations between meridians. Thus, the prediction machine learning models were built up for each meridian in study **I**.

4.1.2. Prediction accuracy of meridians using machine learning approaches

For each of the seven major meridians, we made a comprehensive comparison on the ML model performance on different ML approaches, feature types and data levels. To be specific, we combined four ML approaches (SVM, DT, RF and kNN) with seven types of features (Ext, PubChem, Sub, MACCS, ADME, Ext + ADME and all fingerprints + ADME) in a pairwise manner. Hence, we finally got 84 ML models. The ADME filtering data contained only compounds with higher water solubility and good gastrointestinal absorption (herbs = 583 and compounds = 4,922).

The performance of all the models was evaluated by an accuracy matrix with the 30% dataset after training by 70% dataset. First, we focus on the overall performance of all the 84 models among all the seven meridians. Among all the seven meridians, the Large Intestine meridian achieved the highest overall prediction accuracy at 0.83 on average (p-value < 0.0001, Wilcoxon rank-sum test), followed by the Heart meridian at 0.72 and the Kidney meridian at 0.68 (**I**, Fig 3A). As different organ play different role with compounds, we take into our body. For instance, liver is metabolic while kidney is excretion. Thus, the prediction accuracies for meridians varied among different organs. In addition, the prediction accuracies is dependent of samples size of analysed datasets, with correlation coefficient -0.9. More importantly, Balanced Accuracy and Matthews value of ML models are considerably higher than the corresponding permutation models (**I**, Fig S1, p-value < 0.0001, Wilcoxon rank-sum test), suggesting the general feasibility of predicting the meridians by machine learning models based on the chemical information of herbs and ingredients.

Second, the comparison was performed among three data levels using the Balanced Accuracy metric. It was found that the predictions in compound-level prediction were much better than

those in the herb level (**I**, Fig 3B, p-value < 0.001, Wilcoxon rank-sum test). The superior performance of compound-level predictions compared with the herb-level predictions might be because of the oversimplified way we determined the chemical fingerprint features for an herb. There are three potential reasons for the bad performance of the herb model we built up. First, the ingredients we collected from the TCM database might not be inaccurate in structure because of the limitations of techniques for the extraction and detection of active components in complex herb plants [228]. Second, it is challenging to determine the structural changes of ingredients in the human body caused by the ADME process, especially gut microbiota. The newly generated or lost ingredients would cause changes in therapeutic effects and thus cause bias in the meridian prediction model. Third, for the common ingredients of ingredients, we ignored their differences in actual concentration, which is also one drawback of our study. To sum up, the complex biological roles of the ingredient compounds in herbs vary a lot, such as their concentration, metabolism and inner interactions. However, that information was largely missing from TCMID and other resources. As a result, we assumed that all the ingredient compounds contribute to the pharmacology of the herb equally and thus took the average of its compound features as the herb features. Unlike herb-level data, with average features as the final feature for modelling, the compound-level data consist of compound features and meridians directly, which was more reliable.

As we expected about the importance of ADME properties, filtering out compounds with poor ADME filtering increased the prediction accuracy in the Heart, Lungs and Stomach at the herb level (p-value < 0.05, Wilcoxon rank-sum test), while there is no increased tendency seen for the prediction of Large Intestine and Liver meridians.

To investigate the best machine learning method for each meridian, we compared the prediction performance of different machine learning methods. Based on the finding that the prediction models at the compound level are better than the models at the herb level, the comparison was on models at the compound level. Comparing those top models of seven meridians, we found that RF provides better prediction than kNN, DT and SVM models in general (**I**, Fig 3C), which suggests the advantage of the RF method. To be specific, the RF method was able to detect the predictive features using the Ensemble Learning technique. In other words, RF can avoid overfitting by average accuracy evaluation from multiple-decision trees. At present, many

advanced AI methods, such as Deep Learning, have been applied to the meridian prediction [229]. Neural networks might be promising tools for explaining complex topics, especially in TCM.

The prediction accuracy of machine learning models by different feature types were evaluated. Models with the Ext fingerprint (1024 bits), the longest among all the four fingerprint types, achieved higher accuracy than other features with a p-value <0.05 by Wilcoxon rank-sum test (**I**, Fig 3D). Besides, models using all the fingerprint types combined with ADME performed better than models using them individually in terms of the models with the most accuracies (**I**, Fig 3D). In summary, we demonstrate that the integrated information of the diversity of fingerprints and ADME properties may help to uncover the underlying association for the herb–meridian. Furthermore, the RF method represents a better prediction performance than the other methods.

In addition, the AUROC (the area under the receiver operating characteristic) and the AUPRC (the area under the precision-recall curves) were also calculated and showed a high prediction of machine learning models (**I**, Fig S4). The overall prediction accuracy of the models for each meridian at the compound level by Random Forest using all the available features is listed as a table (**I**, Table 2).

4.1.3. Important fingerprints and ADME features for meridian

After determining the best predictive ability of the meridian classification models with RF as the method and fingerprint and ADME as the features, it is more significant for us to explore the molecular mechanism of the meridian classification. In detail, the importance score value was used to evaluate the contribution of each feature. The importance score was calculated by the change in prediction accuracy at the compound level after a feature was taken away. In detail, provided that dropping a feature causes a sharp decrease in prediction accuracy, this feature will be given a relatively larger score. As we had set up seven models for seven meridians, the top 30 most key features for each meridian were selected, covering 27 ADME properties and 32 fingerprints in total.

To evaluate the importance of the ADME properties across all seven meridians, bio-clustering heatmaps were generated by the importance score of the top ADME feature. As shown in the bi-cluster figure (**I**, Fig 4B), lipophilicity [230], including iLOGP, WLOGP and MLOGP, are top-scored features among all seven meridians, including (mean Z-score 1.66, 0.74 and 0.67, separately). This suggested to us the importance of lipophilicity for the meridian classification. This is consistent with the significant role of Lipophilicity in pharmacokinetic properties and the overall suitability of drug candidates [230]. Another important feature is molar reactivity (MR), with a mean Z-score of 0.96. MR is used to measure the total polarizability of a substance. Additionally, solubility features illustrated comparably higher importance, with Z-around 0.92 to 1.14. In summary, these results indicate that ADME properties are closely related to meridian classification by influencing the pharmacology and pharmacokinetics of ingredients in herb medicine. However, we need to note that some of the 36 ADME properties were determined by prediction models from the SwissADME database. Ideally, the experimentally validated ADME properties would be more accurate and thus help to improve the accuracy. Another limitation is that bioactivity might not stem from the original ingredients in herbs but from the metabolites transformed by gut microbiota or enzymes after oral administration [231]. As a result, more relevant factors that may affect the ADME of herbal medicine deserve to be considered as features to train the models.

We also investigate the top fingerprint features using the bi-cluster method. Interestingly, we found that fingerprint features from the same types tend to cluster together (Rand Index of 0.66) (**I**, Fig 4C). For instance, the most important fingerprint features for the Stomach meridian were clustered together (**I**, Fig 4C). In contrast, the key fingerprints for the Kidney meridian are fingerprints from PubChem. The Spleen and the Lungs are closely related to Ext and MACCS fingerprints, respectively. Generally speaking, Ext fingerprints have the highest score when compared with the other three fingerprint types (**I**, Fig S2). This tendency is also indicated by ML models' better performances by Ext fingerprints (**I**, Fig 3D). To conclude, informative and representative fingerprints are critical for ML prediction.

It is worthy to note that raw TCM herb or other materials might be processed by different methods, such as baking with salt or honey as well as being cut into different parts. The aim of processing is to reduce the contents of toxic constituents, transform the structure of constituents,

or increase the solubility of active constituents. After complex processing, the raw herb was prepared into separate TCM medicine with different therapeutic function or properties (meridian) [232]. In our study, the so-called meridian are based on the prepared TCM medicine and their constituents.

4.2. Network distance models for TCM herb pairs

4.2.1. Frequency of single herbs and herb pairs

In total, 349,197 herb pairs were extracted from 46,929 TCM formulae, covering 4,415 herbs, 4,330 ingredients with target information and 3,171 targets. In the network, 17,753 herb–ingredient pairs and 25,050 ingredient–target associations were utilized for distance calculation among the embedding background network. In terms of the number of herbs contained in one formula, most of the herb formulae (97.9%) consisted of less than 20 herbs (around 4.93 herbs on average). In addition, we found that the top frequency herbs were associated with the immune system. This finding is consistent with the TCM combination principle that tonifying (adjuvant) herbs should be added based on the patient's condition to improve the health of the human body (e.g. immune system), which will help anti-disease [233].

Of the 349,197 herb pairs, herbs of high frequency also have been intensively used in TCM formulae and usually combined with the other herbs (**II,** Fig 2). In addition, 99.4% of the 349,197 herb pairs appeared in no more than 100 TCM formulae, whose combination might be from random chance rather than true synergistic effects and thus not a traditional herb pair. Therefore, the following study only focuses on the top herb pairs for further research as good herb pairs with synergistic effects (frequency ≥ 200).

However, there are some limitations in our study. First, the ingredient concentration is not considered because this kind of information is difficult to mine from the literature or TCM databases due to the inconsistency in different studies. In fact, the actual concentration of ingredients is relevant to the bioactivity effects. In our present methodology, we count the contribution of each ingredient equally, which may influence the accuracy of our model. Furthermore, the second limitation involves the herb pairs we selected. Our study focuses only on the top-frequency herb pairs in a larger number of prescriptions we can extract. There is no

strict threshold for the so-called “good” herb pairs, and not all the synergistic effects of 200 herb pairs were validated by wet experiments. Whereas, it makes sense to some degree, as TCM is an evidence-based methodology; thus, the top frequency herb pairs can be seen as being validated by clinical experience.

4.2.2. Network distance for top-frequent herb pairs

Network modelling was performed at the target level and then at the ingredient level. In total, five distance approaches were tested at each level, including closest, separation, shortest, kernel and center, thus resulting in 25 (5*5) distance models.

The network distance for the top herb 200 was compared to the distance from random herb pairs **(II,** Fig 3). In general, the average network distance of random herb pairs tends to be larger than that of top herb pairs (p-value <0.05). Among all 25 models, 16 show this tendency and the center-separation model showed the largest difference in distance by 0.489 (p-value = 9.91E-28, t-test). As network distance is designed at quantifying the interaction between two subnetworks among the whole PPI background network, the shorter distance of top herb pairs suggests that herb pairs tend to have strong interactions. By these interactions, herb pairs can affect similar pathways to produce synergistic effects.

Similar to our study, the distance-based-mutual-information (DMIM) approach [15] proposed by Li *et al*. is designed to evaluate the interaction score by taking herb frequencies into consideration. Compared with DMIM, our method focused more on molecular-level relationships between herb, ingredient and target. More specifically, with ingredients as the central connection nodes, the interaction between formulae will provide a novel insight for drug combinations that already exist in TCM.

However, our network methods focus on protein–protein interactions, which means that our network model may only help identify combinations due to the MOAs, such as complementary action, neutralizing action and facilitating action but not the pharmacokinetic potentiation. In addition, target information of a majority of compounds, especially natural products, remains unclear although targets of compounds might be relevant to therapeutic effects. The same limitation exists in our study. As targets are the basis of our methodology, herbs or ingredients

were excluded if their target information is missing. To get more compound-target interactions, more advanced computational methods should be developed by considering herbal medicines as holistic system, such as the similarity ensemble approach (SEA)[120], and experimental methods, such as thermal proteomics profiling (TPP) [234].

4.2.3. Performance of the distance metrics for herb combination

Taking the top herb pairs as positive samples and random herb pairs as negative samples, the AUROC and AUPRC were utilized to evaluate the discrimination power of these 25 models. The average AUROC 0.65 and AUPRC 0.72 certified that our network model can characterise the herb–pair interactions **(II,** Table 1). In detail, the center (ingredient)–separation (target) model and the center (ingredient)–shortest (target) model are the top two best models, with both high AUROC and high AUPRC (**II,** Fig 4).

More interestingly, five approaches by the center distance at the ingredient level exhibit a better discrimination performance (mean AUROC 0.80, mean AUPRC 0.83) (**II,** Fig 5) than the other models. The center-based models emphasized the central ingredients. Center ingredients refer to these ingredients that are most close to all the other ingredients in one herb. It hints to us that the central ingredient nodes in one herb may play vital roles for herbal synergistic effects.

To validate our model using external datasets, 268 known herb pairs from the literature were extracted. Finally, we found that top herb pairs are also closer than random herb pairs in the PPI network, suggesting that our model is robust.

4.2.4. MOAs of the herb pair Astragalus membranaceus and Glycyrrhiza uralensis

It has been reported that the combination of *Astragalus membranaceus* and *Glycyrrhiza uralensis* exhibits therapeutic effects for liver diseases [234] by inhibiting bile acid-stimulated inflammation in chronic cholestatic liver injury in mice [223, 224]. However, their active ingredients and MOAs are still unknown. Although oral bioavailability (OB) or drug-likeness (DL) has been widely used for filtering druggable ingredients in TCM studies [33], we did not consider the OB properties and DL properties of ingredients for this work as OB and DL values are usually not from experiments but predictions and might be changed because of the

interactions between the ingredients. Moreover, the OB and DL properties might be changed due to the interactions of the ingredients in the TCM formula.

According to the center-based distance models, the network distances between these two herbs are shorter than those between the top 200 herb pairs in the training set. They are also much smaller than those between the random herb pairs and the top 10,000 top herb pairs (**II**, Table 2).

As we have identified the crucial role of central ingredients, the central ingredients of *Astragalus membranaceus* and *Glycyrrhiza uralensis* were also calculated by our center models. Astramembrannin i and glycyrrhizin were the central ingredients of *Astragalus membranaceus* and *Glycyrrhiza uralensis* respectively. Surprisingly, the synergistic effects for liver diseases of astramembrannin i and glycyrrhizin were supported by the literature [235, 236]. Furthermore, lupeol and isoorientin, by the central ingredients by center (ingredient)–shortest (target) model were also reported to be associated with liver fibrosis disease. For example, isoorientin has protective effects against hepatic fibrosis caused by alcohol through inflammation-related pathways [131]. Additionally, lupeol tends to protect oxidative stress-induced cellular injury of mouse liver by downregulating anti-apoptotic Bcl-2 and upregulating pro-apoptotic Bax and Caspase 3 [237].

In addition to the literature, pathway enrichment and functional analysis were also adopted to uncover the underlying MOAs (**II**, Fig 6). It might be some liver disease-related pathways that contribute to the combination treatment effects of *Astragalus membranaceus* and *Glycyrrhiza uralensis*.

4.3. Network modularity analysis using multipartite network models

In total, 4,485 natural products (herbal medicines) and their 2,857 chemical ingredients from the TCMID database were collected as the two parts of a bipartite network. After getting the projected ISN and NSN network, disconnected nodes were excluded and a multipartite network with 7,004 nodes and 17,555 edges were obtained for further analysis. Using community detection, 42 and 24 communities were prioritized separately using a modularity score from both similarity networks.

We demonstrated that natural products or active ingredients within a cluster tend to have a higher similarity than random clustering (**III**, Fig 4). The network community analysis on ISN and NSN from similar ingredients or natural products indicates that bipartite modeling is useful for the interpretation of TCM. These communities reflect the underlying associations between herbs and ingredients. Specifically, natural products with similar profiles of the ingredients tend to cluster together by community analysis (**III**, Fig 4).

Furthermore, those natural products in the same cluster might also have similar therapeutic effects, which might help to discover novel treatments. Similarly, the cluster of active ingredients in ISN can be used to predict the MOAs of newly discovered or synthesized compounds based on TCM classifications. A functionally unknown molecule with high structural similarity to any of the active ingredient clusters indicates analogous TCM properties and implications. More importantly, as drugs tend to have synergistic effects by acting on distinct biological pathways, members of different clusters might also have different biological effects and thus might be candidates for new drug combinations.

5. CONCLUSION AND FUTURE PERSPECTIVES

Although TCM has attracted extensive interest because of some well-known successful applications, TCM is facing great challenges in gaining recognition from international society due to its unclear constituents, MOAs and toxicity.

Hence, using cutting-edge computational approaches, such as ML and network analysis, we investigated the MOAs of some fundamental TCM concepts, including meridians and formulae, which not only deepen the understanding of TCM but also offer a new viewpoint on how to boost the development of TCM. More importantly, the exploration of TCM would also facilitate drug discovery by identifying the active ingredients from TCM and offering novel disease treatment strategies.

5.1. ML models for meridian prediction

To make full use of valuable TCM resources, the first step is to understand the theories in the TCM system, as they are the fundamental element guiding TCM practice. Hence, in this thesis, we focus on meridian concept which is closely associated with diseases. Until now, the rationale of the meridian for herbs or the human body has yet to be validated by modern methods at the molecular or physical level. AI is emerging as an effective tool for solving many complex problems with the help of the surprising processing ability of computers. In addition, increasingly number of active components of TCM herbs have been identified by researchers for decades. As a result, we proposed a computational framework to investigate meridians with the hypothesis that therapeutic effects of TCM stem from the ingredients in herbs.

First, herb distribution on meridians shows that herbs tend to have multiple meridian classifications, and the modes of distribution also varied among herbs (**II**, Fig 2). In the ML model training, a one-vs-the-rest strategy was applied to each meridian separately. Finally, by embedding the representative feature of herbs with meridian information, the ML models exhibit reliable performance on the meridians' prediction, especially at the compound level (average accuracy = 0.70 in all seven meridians) (**II**, Table 2). These prediction models might be of great significance for herbs or compounds without meridian information.

In short, we brought up novel computational methods to understand the meridian of TCM, and the high prediction accuracy demonstrated that ML is a powerful and promising tool for TCM exploration.

5.2. Meridian theory and molecular properties of ingredients

To ensure the satisfactory performance of ML, the input features matter a lot for ML training. However, the meridian concept only exists for herbs and molecular features are only used to describe ingredients. To get the molecular features of herbs, we merged all the features of ingredients as features of a particular herb. The similar transformation is performed for the meridian of a compound. Those transformations help us to understand the meridian better at the molecular level by integrated relationships (herb–ingredient associations, herb–meridian associations and ingredient–feature associations).

In our ML model, fingerprints were used to represent structural information, and ADME properties were added to represent the absorption, distribution, metabolism and excretion of molecules. The high prediction accuracy of these models illustrated the close associations between molecular features and meridians, especially the compound features with high importance scores.

As a result, ML can not only predict meridians but also be used to interpret each feature's contribution to meridians. The importance score provides us with insightful scopes for the underlying MOAs for meridians. In other words, predictive molecular features might be one factor determining the meridians of herbs and can thus be used for discovering novel active compounds from TCM herbs.

5.3. Network model for quantifying the interactions between TCM formulae

Compared to the one-drug–one-target drug strategy of modern medicine, TCM tends to treat diseases with multiple herbs, multiple ingredients, and multiple targets. To overcome the side effects and drug resistance of single drugs, the drug combination concept was proposed and has become increasingly popular in clinical use, especially for complex diseases. Furthermore, TCM has experienced a long history of herb combination and developed into mature and

comprehensive theories and guidelines. Thus, we believed that formulae would be a valuable resource to discover synergistic ingredients as new drugs. Also, the study on TCM formulae may provide new insight for optimal drug combination strategy.

Despite the potentials of TCM for drug discovery, the MOAs of herb combinations remain unclear to us. Therefore, we exploit the TCM formula as the second study of this thesis. In TCM, herb pairs are considered the foundation of formulae. Consequently, to explore the interactions between herb pairs in TCM, we applied network distance methods to define these interactions.

To be more specific, network distance metrics were performed by integrating herb, ingredients and target information. We found that top herb pairs achieved shorter distances than random herb pairs in all the 25 models. AUROC and AUPRC were used for discrimination evaluation with AUROC 0.65 and AUPRC 0.72 in average. Among all 25 models, the best model was selected for further study, as it has the best discrimination ability for good herb pairs, namely the center (ingredient)–separation (target) model and the center (ingredient)–shortest (target) model.

Taken together, we set up a network model for defining the distance between two herbs in a network, which represents their interactions among PPI networks. With our models, a TCM formula network with herbs as nodes and network distance as weighted length can be further constructed to uncover the hidden associations in TCM.

5.4. Significant role of the central ingredients in the herb pairs

In comparison with the other approaches, the network models we build up focus on the interaction between herbs rather than on single herbs. Using network algorithms, we considered the roles of each ingredient, each herb as well as their associations. For example, by multiple-level networks between herb–ingredients–target–targets, the central distance at the ingredient level achieved better model performance, which suggests the crucial role of central ingredients.

To further validate our finding, we conducted a case study for the herb pair *Astragalus membranaceus* and *Glycyrrhiza uralensis* which have synergistic effects for liver disease. As we expected, the central ingredients lupeol and isoorientin from two herbs have been reported to be relevant to liver fibrosis. The pathway enrichment was also conducted to validate their therapeutic effects for liver diseases.

Overall, through quantified interaction distances at both the ingredient and molecular levels, our network-based model would help to prioritize the potential synergistic ingredients.

5.5. Promise of multipartite network model for TCM study

By community analysis of multipartite networks which were converted from bipartite networks of herbal medicines and their ingredients, we found that herbal medicines among the same clusters tend to share more common properties, meridians, and ingredients from one cluster are more similar in structures and targets. More importantly, multipartite network models can be utilized to better understand the intricate relationship in TCM systematically.

In summary, our study demonstrated that the utilization of computational methods accompanied with multi-omics data would be promising to unlock critical bottlenecks in the understanding of TCM, especially in the aspects of constituents, theories and therapeutic effects. This will be a big step towards TCM modernization.

In the future, it is noteworthy to uncover the active ingredients in herbs, especially their structure and target information. Only in this way can the investigation of TCM be systematic and more accurate. Additionally, in the first study of meridian theory, only four common ML methods were used to probe the relationship between meridian concept and ingredients' features. Other more advanced methods, such as artificial neural networks, could be utilized for a more accurate prediction meridian model as well as models for other TCM concepts. In fact, many advanced machine learning approaches have emerged as interesting candidates for the rationale of TCM, which may considerably accelerate the speed of drug discovery and disease treatment in this field [64]. After deciding the best model for herb pairs, a complex TCM formulae network should be constructed with all the herbs as nodes and their pairwise distance as length weight, which would help uncover the most synergistic herb pairs. Moreover,

more multipartite network models should be generated by integrating a diversity of data in TCM to exploit the hidden relationships in TCM.

ACKNOWLEDGEMENTS

The first project of this dissertation was finished by Jing's group at the Institute for Molecular Medicine Finland (FIMM) and the Oncosys programme of the Faculty of Medicine, University of Helsinki, from 2017.10 to 2021.10. I am grateful to the China Scholarship Council and Academy of Finland Research Fellowship for their financial support in helping me complete my dissertation. I would also like to thank the Doctoral School of Health, Doctoral Programme in Integrative Life Science (ILS), Faculty of Biological and Environmental Sciences and University of Helsinki, for supporting me in attending courses and scientific activities associated with my research field.

My sincere and deepest thanks go to my supervisor Assistant Professor Jing Tang. One of the luckiest things that have ever happened to me is that he accepted me to be your PhD student in 2017. I still remember how ignorant I was at the beginning of my study. It was you who taught me R language step by step and checked my scripts. You give me timely guidance whenever I feel lost in my projects. Your insightful feedback pushed me to sharpen my thinking and brought my work to a higher level. Thanks for all these efforts you put into training me towards my PhD in the past four years. You are one of the most perfect supervisors I have ever seen—patient, insightful, rigorous and friendly.

My sincere gratitude to my second supervisor *Dr.* Mohieddin Jafari for his guidance and inspiration on my projects or daily study. He has broad knowledge and a deep understanding of the network pharmacology and is willing to share his ideas with me. I feel lucky to have him as a collaborator, especially his help in my first and third projects. The common interest on traditional medicine makes us make have wonderful academic communication and collaboration.

My sincere appreciation to Assistant Professor Feixiong Cheng for kindly agreeing to be the opponent for my thesis. I would like to thank *Prof.* Kari P Keinänen for taking time out of his busy schedule to serve as the coordinating academic and the custos. I also would like to express my sincere thanks to my thesis advisory members *Dr.* Henri Xhaard and *Dr.* Jianwei Li for their insightful and constructive suggestions for my projects. I would like to thank the thesis reviewers of my thesis *Prof.* Aik Choon Tan and *Prof.* Feng Zhu for their valuable comments.

I would also like to thank my collaborator *Dr.* Ziaurrehman Tanoli for helping me in a drug repurposing project. Thanks to my communication with him, I learned a lot of skills in multiple public databases extraction and MySQL database operation. The same thanks to Jehad Aldahdooh for his immense help in database construction and server operation.

My sincere gratitude goes to my collaborator *Dr.* Minxia Liu for leading me to the area of precision medicine. She shared a lot of professional knowledge and various innovative analytical approaches about precision medicine with me.

Many thanks to *Dr.* Hongbin Yang for his novel ideas and valuable suggestions on our projects as well the efforts he made to improve them. It is my pleasure to work with him. My great thanks go to Shuyu Zheng, one of the nicest and most helpful people I have ever met, for teaching me a lot, especially in drug combination score calculation, and for her high sense of responsibility. Many thanks to Wenyu Wang for his selfless help in FIMM server operation and genome data analysis. Thanks also to Master student Yingying Hu for her great help for my projects.

My sincere thanks to all my teachers for sharing their knowledge and guiding me through various stages of my education. My special thanks goes to my supervisor *Prof.* Yun Tang in my master's career who led me to the gate of cheminformatics and gave me a lot of guidance and inspiration in network pharmacology analysis.

My deepest thanks go to our group members for the happy time we spent together. Thanks to Alina Malyutina and Joseph Saad for organizing those lovely activities for us. Thanks to Bulat Zagidullin for sharing so many interesting approaches and resources. Thanks to Alberto Pessia and Ali Amiryousefi for the statistical concepts they introduced in group meetings. Thanks to Jie Bao for her time discussing biology with me. Thanks also to Mehdi Mirzaie, Johanna Eriksson, Peter Jakubik, Tolou Shadbahr, Umair Seemab, Cheng Chen and Dalal Aldahdooh for their help in my PhD life.

I would like to thank my dearest Chinese friends at the Meilahti campus: PuChen, Man Xu, Kaiyang Zhang, Yu Fu, Tianduanyi Wang, Liangru Fei, Xiaomeng Xu, Jun Dai, Yafei Wang, Shiqian Li, Kecheng Zhou and Yang Yang. Many appreciations also go to my CSC friends: Yufan Yin, Man Hu, Enpei Zhang, Jian Lv, Changyi Lu, Ran Li, Ming Guo, Zhipeng Tang,

Hao ran Liu, Jian Liu and Binbin Li. My thanks extended to member of climb club: Yixin Liu and Linxiao Chen and members in board game club. Also, I want to express my gratitude to every friend who has appeared in my life for the happy lifetime we have spent together.

Finally, my deepest gratitude to my family members for their great support for my PhD career. (感谢我的家人在整个博士期间对我莫大的支持和帮助). Heartfelt thanks to *Dr.* Yungang He for his accompaniment and encouragement over these years. Although we were separated in two places-China and Finland, our hearts are always connected closer than ever. How lucky we are to have each other's support to survive the whole PhD! Thank you for witnessing my growth from a master to a Doctor. I am looking forward to our happy life in future.

Yinyin Wang

Mar 2021, Helsinki

REFERENCE

1. Antolin, A.A., et al., *Polypharmacology in Precision Oncology: Current Applications and Future Prospects.* Curr Pharm Des, 2016. **22**(46): p. 6935-6945.

2. Tschöp, M.H., et al., *Unimolecular Polypharmacy for Treatment of Diabetes and Obesity.* Cell Metab, 2016. **24**(1): p. 51-62.

3. Reddy, A.S. and S. Zhang, *Polypharmacology: drug discovery for the future.* Expert Rev Clin Pharmacol, 2013. **6**(1): p. 41-7.

4. Hopkins, A.L., *Network pharmacology: the next paradigm in drug discovery.* Nat Chem Biol, 2008. **4**(11): p. 682-90.

5. Silverman, E.K. and J. Loscalzo, *Developing new drug treatments in the era of network medicine.* Clin Pharmacol Ther, 2013. **93**(1): p. 26-8.

6. Bolognesi, M.L. and A. Cavalli, *Multitarget Drug Discovery and Polypharmacology.* ChemMedChem, 2016. **11**(12): p. 1190-2.

7. Cardon, L.R. and J.I. Bell, *Association study designs for complex diseases.* Nat Rev Genet, 2001. **2**(2): p. 91-9.

8. Kibble, M., et al., *Network pharmacology applications to map the unexplored target space and therapeutic potential of natural products.* Nat Prod Rep, 2015. **32**(8): p. 1249-66.

9. Tang, J., et al., *Network pharmacology modeling identifies synergistic Aurora B and ZAK interaction in triple-negative breast cancer.* NPJ Syst Biol Appl, 2019. **5**: p. 20.

10. Malyutina, A., et al., *Drug combination sensitivity scoring facilitates the discovery of synergistic and efficacious drug combinations in cancer.* PLoS Comput Biol, 2019. **15**(5): p. e1006752.

11. Madani Tonekaboni, S.A., et al., *Predictive approaches for drug combination discovery in cancer.* Brief Bioinform, 2018. **19**(2): p. 263-276.

12. Tang, W. and G. Eisenbrand, *Panax ginseng CA Mey*, in *Chinese drugs of plant origin*. 1992, Springer. p. 711-737.

13. Wang, S., et al., *Compatibility art of traditional Chinese medicine: from the perspective of herb pairs.* J Ethnopharmacol, 2012. **143**(2): p. 412-23.

14. Zhou, M., et al., *Recent pharmaceutical evidence on the compatibility rationality of traditional Chinese medicine.* J Ethnopharmacol, 2017. **206**: p. 363-375.

15. Li, S., et al., *Herb network construction and co-module analysis for uncovering the combination rule of traditional Chinese herbal formulae.* BMC Bioinformatics, 2010. **11 Suppl 11**(Suppl 11): p. S6.

16. Zhao, X., et al., *A novel drug discovery strategy inspired by traditional medicine philosophies.* Science, 2015. **347**(6219): p. S38-S40.

17. Chan, K., *Progress in traditional Chinese medicine.* Trends Pharmacol Sci, 1995. **16**(6): p. 182-7.

18. Gu, S. and J. Pei, *Innovating Chinese Herbal Medicine: From Traditional Health Practice to Scientific Drug Discovery.* Front Pharmacol, 2017. **8**: p. 381.

19. Rezadoost, H., M. Karimi, and M. Jafari, *Proteomics of hot-wet and cold-dry temperaments proposed in Iranian traditional medicine: a Network-based Study.* Sci Rep, 2016. **6**: p. 30133.

20. Jafari, M., et al., *Proteomics and traditional medicine: new aspect in explanation of temperaments.* Forsch Komplementmed, 2014. **21**(4): p. 250-3.

21. Chon, T.Y. and M.C. Lee, *Acupuncture.* Mayo Clin Proc, 2013. **88**(10): p. 1141-6.

22. Azizkhani, M., et al., *Traditional dry cupping therapy versus medroxyprogesterone acetate in the treatment of idiopathic menorrhagia: a randomized controlled trial.* 2018. **20**(2).

23. Wang, G.J., M.H. Ayati, and W.B. Zhang, *Meridian studies in China: a systematic review.* J Acupunct Meridian Stud, 2010. **3**(1): p. 1-9.

24. Liang, H., et al., *Herb-target interaction network analysis helps to disclose molecular mechanism of traditional Chinese medicine.* Sci Rep, 2016. **6**: p. 36767.

25. Zhang, C., et al., *Deciphering Potential Correlations between New Biomarkers and Pattern Classification in Chinese Medicine by Bioinformatics: Two Examples of Rheumatoid Arthritis.* Chin J Integr Med, 2018.

26. Bai, Y., et al., *Review of evidence suggesting that the fascia network could be the anatomical basis for acupoints and meridians in the human body.* Evid Based Complement Alternat Med, 2011. **2011**: p. 260510.

27. Ma, W., et al., *Perivascular space: possible anatomical substrate for the meridian.* J Altern Complement Med, 2003. **9**(6): p. 851-9.

28. Jie, Z., et al., *General Medication Rules in Treating Spleen-stomach Disharmony Based on Traditional Chinese Medicine Inheritance Platform.* World Chinese Medicine, 2016. **1**: p. 048.

29. Lin, F.Y., et al., *Controversial opinion: evaluation of EGR1 and LAMA2 loci for high myopia in Chinese populations.* J Zhejiang Univ Sci B, 2016. **17**(3): p. 225-35.

30. Fu, X.J., et al., *A study on the antioxidant activity and tissues selective inhibition of lipid peroxidation by saponins from the roots of Platycodon grandiflorum.* Am J Chin Med, 2009. **37**(5): p. 967-75.

31. Wang, X., S.L. Morris-Natschke, and K.H. Lee, *New developments in the chemistry and biology of the bioactive constituents of Tanshen.* Med Res Rev, 2007. **27**(1): p. 133-48.

32. Li, Z.M., S.W. Xu, and P.Q. Liu, *Salvia miltiorrhizaBurge (Danshen): a golden herbal medicine in cardiovascular therapeutics.* Acta Pharmacol Sin, 2018. **39**(5): p. 802-824.

33. Huang, C., et al., *Systems pharmacology in drug discovery and therapeutic insight for herbal medicines.* Brief Bioinform, 2014. **15**(5): p. 710-33.

34. Yang, X., et al., *Information integration research on cumulative effect of 'Siqi, Wuwei, and Guijing' in Traditional Chinese Medicine.* J Tradit Chin Med, 2016. **36**(4): p. 538-46.

35. Cheng, F., I.A. Kovács, and A.L. Barabási, *Network-based prediction of drug combinations.* Nat Commun, 2019. **10**(1): p. 1197.

36. Zhang, R., et al., *Machine learning approaches for elucidating the biological effects of natural products.* Nat Prod Rep, 2021. **38**(2): p. 346-361.

37. Fang, J., et al., *In silico polypharmacology of natural products.* Brief Bioinform, 2018. **19**(6): p. 1153-1171.

38. Fu, X., et al., *Toward Understanding the Cold, Hot, and Neutral Nature of Chinese Medicines Using in Silico Mode-of-Action Analysis.* J Chem Inf Model, 2017. **57**(3): p. 468-483.

39. Wang, M., et al., *Classification of Mixtures of Chinese Herbal Medicines Based on a Self-organizing Map (SOM).* Mol Inform, 2016. **35**(3-4): p. 109-15.

40. Zhou, W., et al., *Systems pharmacology exploration of botanic drug pairs reveals the mechanism for treating different diseases.* Sci Rep, 2016. **6**: p. 36985.

41. Ung, C.Y., et al., *Are herb-pairs of traditional Chinese medicine distinguishable from others? Pattern analysis and artificial intelligence classification study of traditionally defined herbal properties.* J Ethnopharmacol, 2007. **111**(2): p. 371-7.

42. Cao, J., *The Common Prescription Patterns Based on the Hierarchical Clustering of Herb-Pairs Efficacies.* Evid Based Complement Alternat Med, 2016. **2016**: p. 6373270.

43. Yang, M., et al., *Application of genetic algorithm for discovery of core effective formulae in TCM clinical data.* Comput Math Methods Med, 2013. **2013**: p. 971272.

44. Tang, J. and T. Aittokallio, *Network pharmacology strategies toward multi-target anticancer therapies: from computational models to experimental design principles.* Curr Pharm Des, 2014. **20**(1): p. 23-36.

45. Li, S. and B. Zhang, *Traditional Chinese medicine network pharmacology: theory, methodology and application.* Chin J Nat Med, 2013. **11**(2): p. 110-20.

46. Li, S., et al., *Network pharmacology in traditional chinese medicine.* Evid Based Complement Alternat Med, 2014. **2014**: p. 138460.

47. Li, J., et al., *Traditional chinese medicine-based network pharmacology could lead to new multicompound drug discovery.* Evid Based Complement Alternat Med, 2012. **2012**: p. 149762.

48. Stone, R., *Biochemistry. Lifting the veil on traditional Chinese medicine.* Science, 2008. **319**(5864): p. 709-10.

49. Xu, Z., *Modernization: One step at a time.* Nature, 2011. **480**(7378): p. S90-2.

50. Kim, H.U., et al., *A systems approach to traditional oriental medicine.* Nat Biotechnol, 2015. **33**(3): p. 264-8.

51. Li, H., et al., *Cardioprotective effect of paeonol and danshensu combination on isoproterenol-induced myocardial injury in rats.* PLoS One, 2012. **7**(11): p. e48872.

52. Xue, L., et al., *Effects and interaction of icariin, curculigoside, and berberine in er-xian decoction, a traditional chinese medicinal formula, on osteoclastic bone resorption.* Evid Based Complement Alternat Med, 2012. **2012**: p. 490843.

53. Saw, C.L., et al., *Pharmacodynamics of ginsenosides: antioxidant activities, activation of Nrf2, and potential synergistic effects of combinations.* Chem Res Toxicol, 2012. **25**(8): p. 1574-80.

54. Yuan, H., et al., *The Traditional Medicine and Modern Medicine from Natural Products.* Molecules, 2016. **21**(5).

55. Fortunato, S., *Community detection in graphs.* Phys Rep, 2010. **486**(3): p. 75-174.

56. Weston, S., et al., *Broad Anti-coronavirus Activity of Food and Drug Administration-Approved Drugs against SARS-CoV-2 In Vitro and SARS-CoV In Vivo.* J Virol, 2020. **94**(21).

57. Girvan, M. and M.E. Newman, *Community structure in social and biological networks.* Proc Natl Acad Sci U S A, 2002. **99**(12): p. 7821-6.

58. Junker, B.H. and F. Schreiber, *Analysis of Biological Networks.* Wiley Online Library. (John Wiley & Sons, Inc.), 2008: p. 3-12.

59. Agnarsson, G. and R. Greenlaw, *Graph Theory: Modeling.* Applications, and Algorithms, 2007.

60. Samuelson, G. and L. Bohlin, *Drugs of Natural origin.* A texboot of Pharmacognosy. Stockholm: Swedish Pharmaceutical Pres, 1992.

61. Hanson, J.R., *Natural products: the secondary metabolites*. Vol. 17. 2003: Royal Society of Chemistry.

62. Brahmachari, G., *Bioactive Natural Products*. 2015: Wiley Online Library.

63. Cushnie, T.P.T., et al., *Bioprospecting for Antibacterial Drugs: a Multidisciplinary Perspective on Natural Product Source Material, Bioassay Selection and Avoidable Pitfalls.* Pharm Res, 2020. **37**(7): p. 125.

64. Rodrigues, T., et al., *Counting on natural products for drug design.* Nat Chem, 2016. **8**(6): p. 531-41.

65. Chen, Y., C. de Bruyn Kops, and J. Kirchmair, *Data Resources for the Computer-Guided Discovery of Bioactive Natural Products.* J Chem Inf Model, 2017. **57**(9): p. 2099-2111.

66. Dias, D.A., S. Urban, and U. Roessner, *A historical overview of natural products in drug discovery.* Metabolites, 2012. **2**(2): p. 303-36.

67. Schrör, K., *Acetylsalicylic acid*. 2016: John Wiley & Sons.

68. World Health, O., *WHO traditional medicine strategy: 2014-2023*. 2013: World Health Organization.

69. Curd, M. and J.A. Cover, *Philosophy of science: The central issues.* 1998.

70. Sneader, W., *Drug discovery: a history*. 2005: John Wiley & Sons.

71. Sampson, W., *Antiscience trends in the rise of the "alternative medicine" movement.* Ann N Y Acad Sci, 1996. **775**: p. 188-97.

72. Cragg, G.M. and D.J. Newman, *Natural products: A continuing source of novel drug leads.* Biophys Acta Gen Subj, 2013. **1830**(6): p. 3670-3695.

73. Tu, Y., *The discovery of artemisinin (qinghaosu) and gifts from Chinese medicine.* Nat Med, 2011. **17**(10): p. 1217-20.

74. Shen, Z.X., et al., *Use of arsenic trioxide (As2O3) in the treatment of acute promyelocytic leukemia (APL): II. Clinical efficacy and pharmacokinetics in relapsed patients.* Blood, 1997. **89**(9): p. 3354-60.

75. Yan, Q., *Stress and Systemic Inflammation: Yin-Yang Dynamics in Health and Diseases.* Methods Mol Biol, 2018. **1781**: p. 3-20.

76. Fei, X., et al., *Probing the Qi of traditional Chinese herbal medicines by the biological synthesis of nano-Au.* J Mater Chem B, 2018. **6**(19): p. 3156-3162.

77. Zhang, W.B., G.J. Wang, and K. Fuxe, *Classic and Modern Meridian Studies: A Review of Low Hydraulic Resistance Channels along Meridians and Their Relevance for Therapeutic Effects in Traditional Chinese Medicine.* Evid Based Complement Alternat Med, 2015. **2015**: p. 410979.

78. Longhurst, J.C., *Defining meridians: a modern basis of understanding.* J Acupunct Meridian Stud, 2010. **3**(2): p. 67-74.

79. Wang, W.J. and T. Zhang, *Integration of traditional Chinese medicine and Western medicine in the era of precision medicine.* J Integr Med, 2017. **15**(1): p. 1-7.

80. Hsieh, H.Y., P.H. Chiu, and S.C. Wang, *Epigenetics in traditional chinese pharmacy: a bioinformatic study at pharmacopoeia scale.* Evid Based Complement Alternat Med, 2011. **2011**: p. 816714.

81. Jiang, M., et al., *Syndrome differentiation in modern research of traditional Chinese medicine.* J Ethnopharmacol, 2012. **140**(3): p. 634-42.

82. Zhang, B., X. Wang, and S. Li, *An Integrative Platform of TCM Network Pharmacology and Its Application on a Herbal Formula, Qing-Luo-Yin.* Evid Based Complement Alternat Med, 2013. **2013**: p. 456747.

83. Ma, T., et al., *Bridging the gap between traditional Chinese medicine and systems biology: the connection of Cold Syndrome and NEI network.* Mol Biosyst, 2010. **6**(4): p. 613-9.

84. Li, S., et al., *Understanding ZHENG in traditional Chinese medicine in the context of neuro-endocrine-immune network.* IET Syst Biol, 2007. **1**(1): p. 51-60.

85. Fung, F.Y. and Y.C. Linn, *Developing traditional chinese medicine in the era of evidence-based medicine: current evidences and challenges.* Evid Based Complement Alternat Med, 2015. **2015**: p. 425037.

86. Zhang, R., et al., *Network Pharmacology Databases for Traditional Chinese Medicine: Review and Assessment.* Front Pharmacol, 2019. **10**: p. 123.

87. Sánchez-Vidaña, D.I., R. Rajwani, and M.S. Wong, *The Use of Omic Technologies Applied to Traditional Chinese Medicine Research.* Evid Based Complement Alternat Med, 2017. **2017**: p. 6359730.

88. Chen, X., et al., *Database of traditional Chinese medicine and its application to studies of mechanism and to prescription validation.* Br J Pharmacol, 2006. **149**(8): p. 1092-103.

89. Chen, C.Y., *TCM Database@Taiwan: the world's largest traditional Chinese medicine database for drug screening in silico.* PLoS One, 2011. **6**(1): p. e15939.

90. Ru, J., et al., *TCMSP: a database of systems pharmacology for drug discovery from herbal medicines.* J Cheminform, 2014. **6**: p. 13.

91. Huang, L., et al., *TCMID 2.0: a comprehensive resource for TCM.* Nucleic Acids Res, 2018. **46**(D1): p. D1117-d1120.

92. Ye, H., et al., *HIT: linking herbal active ingredients to targets.* Nucleic Acids Res, 2011. **39**(Database issue): p. D1055-9.

93. Zhang, R.Z., et al., *TCM-Mesh: The database and analytical system for network pharmacology analysis for TCM preparations.* Sci Rep, 2017. **7**(1): p. 2821.

94. Kuhn, M., et al., *The SIDER database of drugs and side effects.* Nucleic Acids Res, 2016. **44**(D1): p. D1075-9.

95. Li, B., et al., *YaTCM: Yet another Traditional Chinese Medicine Database for Drug Discovery.* Comput Struct Biotechnol J, 2018. **16**: p. 600-610.

96. Xu, H.Y., et al., *ETCM: an encyclopaedia of traditional Chinese medicine.* Nucleic Acids Res, 2019. **47**(D1): p. D976-d982.

97. Liu, Z., et al., *TCMAnalyzer: A Chemo- and Bioinformatics Web Service for Analyzing Traditional Chinese Medicine.* J Chem Inf Model, 2018. **58**(3): p. 550-555.

98. Liu, Z., et al., *BATMAN-TCM: a Bioinformatics Analysis Tool for Molecular mechANism of Traditional Chinese Medicine.* Sci Rep, 2016. **6**: p. 21146.

99. Kim, S.K., et al., *TM-MC: a database of medicinal materials and chemical compounds in Northeast Asian traditional medicine.* BMC Complement Altern Med, 2015. **15**: p. 218.

100. Wu, Y., et al., *SymMap: an integrative database of traditional Chinese medicine enhanced by symptom mapping.* Nucleic Acids Res, 2019. **47**(D1): p. D1110-d1117.

101. Fang, S., et al., *HERB: a high-throughput experiment- and reference-guided database of traditional Chinese medicine.* Nucleic Acids Res, 2021. **49**(D1): p. D1197-d1206.

102. Liu, Z., et al., *TCMIO: A Comprehensive Database of Traditional Chinese Medicine on Immuno-Oncology.* Front Pharmacol, 2020. **11**: p. 439.

103. Mirzaeian, R., et al., *Progresses and challenges in the traditional medicine information system: A systematic review.* Journal of Pharmacy & Pharmacognosy Research, 2019.

104. Lukman, S., Y. He, and S.C. Hui, *Computational methods for Traditional Chinese Medicine: a survey.* Comput Methods Programs Biomed, 2007. **88**(3): p. 283-94.

105. Zhou, X., et al., *Development of traditional Chinese medicine clinical data warehouse for medical knowledge discovery and decision support.* Artif Intell Med, 2010. **48**(2-3): p. 139-52.

106. Cao, C., H. Wang, and Y. Sui, *Knowledge modeling and acquisition of traditional Chinese herbal drugs and formulae from text.* Artif Intell Med, 2004. **32**(1): p. 3-13.

107. Li, C., et al. *TCMiner: A High Performance Data Mining System for Multi-dimensional Data Analysis of Traditional Chinese Medicine Prescriptions*. in *Conceptual Modeling*

for Advanced Application Domains. 2004. Berlin, Heidelberg: Springer Berlin Heidelberg.

108. Zhou, J., G. Xie, and X.J.I.C.A. Yan, *Encyclopedia of traditional Chinese medicines.* 2011. **1**: p. 455.

109. Zhou, X., et al., *Data mining in real-world traditional Chinese medicine clinical data warehouse*, in *Data Analytics for Traditional Chinese Medicine Research*. 2014, Springer. p. 189-213.

110. Zhou, X., et al., *Ontology development for unified traditional Chinese medical language system.* Artif Intell Med, 2004. **32**(1): p. 15-27.

111. Shao, L. and B. Zhang, *Traditional Chinese medicine network pharmacology: theory, methodology and application.* Chin J Nat Med, 2013. **11**(2): p. 110-120.

112. Gaulton, A., et al., *ChEMBL: a large-scale bioactivity database for drug discovery.* Nucleic Acids Res, 2012. **40**(Database issue): p. D1100-7.

113. Szklarczyk, D., et al., *STITCH 5: augmenting protein-chemical interaction networks with tissue and affinity data.* Nucleic Acids Res, 2016. **44**(D1): p. D380-4.

114. Chen, Y.Z. and D.G. Zhi, *Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule.* Proteins, 2001. **43**(2): p. 217-26.

115. Berman, H.M., et al., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42.

116. Yu, H., et al., *A systematic prediction of multiple drug-target interactions from chemical, genomic, and pharmacological data.* PLoS One, 2012. **7**(5): p. e37608.

117. Wang, Z., et al., *Improving chemical similarity ensemble approach in target prediction.* J Cheminform, 2016. **8**: p. 20.

118. Wu, Z., et al., *In silico prediction of chemical mechanism of action via an improved network-based inference method.* Br J Pharmacol, 2016. **173**(23): p. 3372-3385.

119. Jung, S.H., *Stratified Fisher's exact test and its sample size calculation.* Biom J, 2014. **56**(1): p. 129-40.

120. Gu, S. and L.-h. Lai, *Associating 197 Chinese herbal medicine with drug targets and diseases using the similarity ensemble approach.* Acta Pharmacol Sin, 2020. **41**(3): p. 432-438.

121. Yang, X., et al., *[New-generation high-throughput technologies based 'omics' research strategy in human disease].* Yi Chuan, 2011. **33**(8): p. 829-46.

122. Del Giacco, L. and C. Cattaneo, *Introduction to genomics.* Methods Mol Biol, 2012. **823**: p. 79-88.

123. Kang, Y.J., *Herbogenomics: from traditional Chinese medicine to novel therapeutics.* Exp Biol Med (Maywood), 2008. **233**(9): p. 1059-65.

124. Ling, C.Q., et al., *The roles of traditional Chinese medicine in gene therapy.* J Integr Med, 2014. **12**(2): p. 67-75.

125. Lv, C., et al., *The gene expression profiles in response to 102 traditional Chinese medicine (TCM) components: a general template for research on TCMs.* Sci Rep, 2017. **7**(1): p. 352.

126. Liu, M., et al., *Transcriptional profiling of Chinese medicinal formula Si-Wu-Tang on breast cancer cells reveals phytoestrogenic activity.* BMC Complement Altern Med, 2013. **13**: p. 11.

127. Yoo, M., et al., *Exploring the molecular mechanisms of Traditional Chinese Medicine components using gene expression signatures and connectivity map.* Comput Methods Programs Biomed, 2019. **174**: p. 33-40.

128. Clough, E. and T. Barrett, *The Gene Expression Omnibus Database.* Methods Mol Biol, 2016. **1418**: p. 93-110.

129. Gika, H.G., G.A. Theodoridis, and I.D. Wilson, *Metabolic Profiling: Status, Challenges, and Perspective.* Methods Mol Biol, 2018. **1738**: p. 3-13.

130. Han, Y., et al., *Chinmedomics, a new strategy for evaluating the therapeutic efficacy of herbal medicines.* Pharmacol Ther, 2020. **216**: p. 107680.

131. Huang, Q.F., et al., *Protective effect of isoorientin-2″-O-α-L-arabinopyranosyl isolated from Gypsophila elegans on alcohol induced hepatic fibrosis in rats.* Food Chem Toxicol, 2012. **50**(6): p. 1992-2001.

132. Wang, X., A. Zhang, and H. Sun, *Future perspectives of Chinese medical formulae: chinmedomics as an effector.* Omics, 2012. **16**(7-8): p. 414-21.

133. Feng, W., et al., *Gut microbiota, a new frontier to understand traditional Chinese medicines.* Pharmacol Res, 2019. **142**: p. 176-191.

134. Chang, C.J., et al., *Corrigendum: Ganoderma lucidum reduces obesity in mice by modulating the composition of the gut microbiota.* Nat Commun, 2017. **8**: p. 16130.

135. Nie, Q., et al., *Dietary compounds and traditional Chinese medicine ameliorate type 2 diabetes by modulating gut microbiota.* Crit Rev Food Sci Nutr, 2019. **59**(6): p. 848-863.

136. Wang, X., et al., *Urine metabolomics analysis for biomarker discovery and detection of jaundice syndrome in patients with liver disease.* Mol Cell Proteomics, 2012. **11**(8): p. 370-80.

137. Jiang, B., et al., *Integrating next-generation sequencing and traditional tongue diagnosis to determine tongue coating microbiome.* Sci Rep, 2012. **2**: p. 936.

138. Moffat, J.G., et al., *Opportunities and challenges in phenotypic drug discovery: an industry perspective.* Nat Rev Drug Discov, 2017. **16**(8): p. 531-543.

139. Poole, D., A. Mackworth, and R. Goebel, *Computational intelligence: a logical approach.(1998).* Google Scholar Google Scholar Digital Library Digital Library, 1998.

140. Zhu, H., *Big Data and Artificial Intelligence Modeling for Drug Discovery.* Annu Rev Pharmacol Toxicol, 2020. **60**: p. 573-589.

141. Hosny, A., et al., *Artificial intelligence in radiology.* Nat Rev Cancer, 2018. **18**(8): p. 500-510.

142. Bishop, C.M., *Pattern recognition and machine learning*. 2006: springer.

143. Russell, S. and P. Norvig, *Artificial intelligence: a modern approach.* 2002.

144. Wang, X., *A Fast Exact k-Nearest Neighbors Algorithm for High Dimensional Search Using k-Means Clustering and Triangle Inequality.* Proc Int Jt Conf Neural Netw, 2012. **43**(6): p. 2351-2358.

145. Tahiri, N., M. Willems, and V. Makarenkov, *A new fast method for inferring multiple consensus trees using k-medoids.* BMC Evol Biol, 2018. **18**(1): p. 48.

146. Bezdek, J.C., R. Ehrlich, and W. Full, *FCM: The fuzzy c-means clustering algorithm.* Computers & geosciences, 1984. **10**(2-3): p. 191-203.

147. Buckley, J.J., *Fuzzy hierarchical analysis.* Fuzzy sets and systems, 1985. **17**(3): p. 233-247.

148. Cunningham, P., M. Cord, and S.J. Delany, *Supervised learning*, in *Machine learning techniques for multimedia*. 2008, Springer. p. 21-49.

149. Ramsey, J.B.J.J.o.t.R.S.S.S.B., *Tests for specification errors in classical linear least‐squares regression analysis.* 1969. **31**(2): p. 350-371.

150. Ruczinski, I., C. Kooperberg, and M. LeBlanc, *Logic regression.* J Comput Graph Stat, 2003. **12**(3): p. 475-511.

151. Solberg, H.E., *Discriminant analysis.* CRC Crit Rev Clin Lab Sci, 1978. **9**(3): p. 209-42.

152. Bergh, D.V.D., et al., *A tutorial on Bayesian multi-model linear regression with BAS and JASP.* Behav Res Methods, 2021.

153. Guo, G., et al. *KNN Model-Based Approach in Classification*. in *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*. 2003. Berlin, Heidelberg: Springer Berlin Heidelberg.

154. Haifeng, W. and H. Dejin. *Comparison of SVM and LS-SVM for Regression*. in *2005 International Conference on Neural Networks and Brain*. 2005.

155. Kingsford, C. and S.L. Salzberg, *What are decision trees?* Nature Biotechnology, 2008. **26**(9): p. 1011-1013.

156. Svetnik, V., et al., *Random forest: a classification and regression tool for compound classification and QSAR modeling.* J Chem Inf Comput Sci, 2003. **43**(6): p. 1947-58.

157. Krogh, A., *What are artificial neural networks?* Nat Biotechnol, 2008. **26**(2): p. 195-7.

158. Rodríguez-Rodríguez, I., et al., *Applications of Artificial Intelligence, Machine Learning, Big Data and the Internet of Things to the COVID-19 Pandemic: A Scientometric Review Using Text Mining.* Int J Environ Res Public Health, 2021. **18**(16).

159. Cangelosi, D., et al., *Artificial neural network classifier predicts neuroblastoma patients' outcome.* BMC Bioinformatics, 2016. **17**(Suppl 12): p. 347.

160. Kourou, K., et al., *Machine learning applications in cancer prognosis and prediction.* Comput Struct Biotechnol J, 2015. **13**: p. 8-17.

161. Senior, A.W., et al., *Improved protein structure prediction using potentials from deep learning.* Nature, 2020. **577**(7792): p. 706-710.

162. Liu, P., et al., *Deep Evolutionary Networks with Expedited Genetic Algorithms for Medical Image Denoising.* Med Image Anal, 2019. **54**: p. 306-315.

163. Mohamed, A., C.H. Nguyen, and H. Mamitsuka, *Current status and prospects of computational resources for natural product dereplication: a review.* Brief Bioinform, 2016. **17**(2): p. 309-21.

164. Arji, G., et al., *A systematic literature review and classification of knowledge discovery in traditional medicine.* Comput Methods Programs Biomed, 2019. **168**: p. 39-57.

165. Yu, T., et al., *Knowledge graph for TCM health preservation: Design, construction, and applications.* Artif Intell Med, 2017. **77**: p. 48-52.

166. Anastasi, J.K., L.M. Currie, and G.H. Kim, *Understanding diagnostic reasoning in TCM practice: tongue diagnosis.* Altern Ther Health Med, 2009. **15**(3): p. 18-28.

167. Shu, J.J. and Y. Sun, *Developing classification indices for Chinese pulse diagnosis.* Complement Ther Med, 2007. **15**(3): p. 190-8.

168. Wang, H. and Y. Cheng, *A quantitative system for pulse diagnosis in Traditional Chinese Medicine.* Conf Proc IEEE Eng Med Biol Soc, 2005. **2005**: p. 5676-9.

169. Pang, B., et al., *Computerized tongue diagnosis based on Bayesian networks.* IEEE Trans Biomed Eng, 2004. **51**(10): p. 1803-10.

170. Su, S.B., et al., *Evidence-Based ZHENG: A Traditional Chinese Medicine Syndrome.* Evid Based Complement Alternat Med, 2012. **2012**: p. 246538.

171. Hu, Y., et al., *Automatic Construction of Chinese Herbal Prescriptions From Tongue Images Using CNNs and Auxiliary Latent Therapy Topics.* IEEE Trans Cybern, 2021. **51**(2): p. 708-721.

172. Zhang, N.L., et al., *Latent tree models and diagnosis in traditional Chinese medicine.* Artif Intell Med, 2008. **42**(3): p. 229-45.

173. Zhao, Y., et al. *TCM syndrome differentiation of AIDS using subspace clustering algorithm*. in *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2014.

174. Ruan, C., et al. *THCluster: herb supplements categorization for precision traditional Chinese medicine*. in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2017. IEEE.

175. Chen, X., M.X. Liu, and G.Y. Yan, *Drug-target interaction prediction by random walk on the heterogeneous network.* Mol Biosyst, 2012. **8**(7): p. 1970-8.

176. Maetschke, S.R., et al., *Supervised, semi-supervised and unsupervised inference of gene regulatory networks.* Brief Bioinform, 2014. **15**(2): p. 195-211.

177. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.

178. Harris, M.A., et al., *The Gene Ontology (GO) database and informatics resource.* Nucleic Acids Res, 2004. **32**(Database issue): p. D258-61.

179. Vanunu, O., et al., *Associating genes and protein complexes with disease via network propagation.* PLoS Comput Biol, 2010. **6**(1): p. e1000641.

180. Xie, G., et al., *Poly-pharmacokinetic Study of a Multicomponent Herbal Medicine in Healthy Chinese Volunteers.* Clin Pharmacol Ther, 2018. **103**(4): p. 692-702.

181. Lu, C., et al., *Network-based gene expression biomarkers for cold and heat patterns of rheumatoid arthritis in traditional chinese medicine.* Evid Based Complement Alternat Med, 2012. **2012**: p. 203043.

182. Ramsay, R.R., et al., *A perspective on multi-target drug discovery and design for complex diseases.* Clin Transl Med, 2018. **7**(1): p. 3.

183. Qiu, J., *Traditional medicine: a culture in the balance.* Nature, 2007. **448**(7150): p. 126-8.

184. Huang, L., et al., *DrugComboRanker: drug combination discovery based on target network analysis.* Bioinformatics, 2014. **30**(12): p. i228-36.

185. Jafari, M., et al., *Unsupervised Learning and Multipartite Network Models: A Promising Approach for Understanding Traditional Medicine.* Front Pharmacol, 2020. **11**: p. 1319.

186. Yue, S.J., et al., *Herb pair Danggui-Honghua: mechanisms underlying blood stasis syndrome by system pharmacology approach.* Sci Rep, 2017. **7**: p. 40318.

187. Zhao, F.R., et al., *Antagonistic effects of two herbs in Zuojin Wan, a traditional Chinese medicine formula, on catecholamine secretion in bovine adrenal medullary cells.* Phytomedicine, 2010. **17**(8-9): p. 659-68.

188. Chen, Y.F., et al., *Influence of Zuojin pill and retro-Zuojin pill on inflammatory and protective factors in rats with gastric mucosa lesion of cold and hot type.* Chin J Integr Tradit West Med Dig, 2003. **11**(3): p. 133-135.

189. Hu, Y., et al., *Inhibitory effect and transcriptional impact of berberine and evodiamine on human white preadipocyte differentiation.* Fitoterapia, 2010. **81**(4): p. 259-68.

190. Liu, S.H., et al., *Safety surveillance of traditional Chinese medicine: current and future.* Drug Saf, 2015. **38**(2): p. 117-28.

191. Morschheuser, L., et al., *High-performance thin-layer chromatography as a fast screening tool for phosphorylated peptides.* J Chromatogr B Analyt Technol Biomed Life Sci, 2016. **1008**: p. 198-205.

192. Meyer, V.R., *Practical high-performance liquid chromatography*. 2013: John Wiley & Sons.

193. Teschke, R., et al., *Traditional Chinese Medicine (TCM) and Herbal Hepatotoxicity: RUCAM and the Role of Novel Diagnostic Biomarkers Such as MicroRNAs.* Medicines (Basel), 2016. **3**(3).

194. Newman, D.J., *Modern traditional Chinese medicine: Identifying, defining and usage of TCM components.* Adv Pharmacol, 2020. **87**: p. 113-158.

195. Lin, T.L., et al., *Role of gut microbiota in identification of novel TCM-derived active metabolites.* Protein Cell, 2020.

196. Daina, A., O. Michielin, and V. Zoete, *SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules.* Sci Rep, 2017. **7**: p. 42717.

197. Licata, L., et al., *MINT, the molecular interaction database: 2012 update.* Nucleic Acids Res, 2012. **40**(Database issue): p. D857-61.

198. Orchard, S., et al., *The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases.* Nucleic Acids Res, 2014. **42**(Database issue): p. D358-63.

199. Breuer, K., et al., *InnateDB: systems biology of innate immunity and beyond--recent updates and continuing curation.* Nucleic Acids Res, 2013. **41**(Database issue): p. D1228-33.

200. Cowley, M.J., et al., *PINA v2.0: mining interactome modules.* Nucleic Acids Res, 2012. **40**(Database issue): p. D862-5.

201. Peri, S., et al., *Human protein reference database as a discovery resource for proteomics.* Nucleic Acids Res, 2004. **32**(Database issue): p. D497-501.

202. Rolland, T., et al., *A proteome-scale map of the human interactome network.* Cell, 2014. **159**(5): p. 1212-1226.

203. Hornbeck, P.V., et al., *PhosphoSitePlus, 2014: mutations, PTMs and recalibrations.* Nucleic Acids Res, 2015. **43**(Database issue): p. D512-20.

204. Cheng, F., et al., *Quantitative network mapping of the human kinome interactome reveals new clues for rational kinase inhibitor discovery and individualized cancer therapy.* Oncotarget, 2014. **5**(11): p. 3697-710.

205. Meyer, M.J., et al., *INstruct: a database of high-quality 3D structurally resolved protein interactome networks.* Bioinformatics, 2013. **29**(12): p. 1577-9.

206. Fazekas, D., et al., *SignaLink 2 - a signaling pathway resource with multi-layered regulatory networks.* BMC Syst Biol, 2013. **7**: p. 7.

207. Chatr-Aryamontri, A., et al., *The BioGRID interaction database: 2015 update.* Nucleic Acids Res, 2015. **43**(Database issue): p. D470-8.

208. Arús-Pous, J., et al., *Randomized SMILES strings improve the quality of molecular generative models.* J Cheminform, 2019. **11**(1): p. 71.

209. O'Boyle, N.M., et al., *Open Babel: An open chemical toolbox.* J Cheminform, 2011. **3**: p. 33.

210. Yap, C.W., *PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints.* J Comput Chem, 2011. **32**(7): p. 1466-74.

211. Rogers, D. and M. Hahn, *Extended-connectivity fingerprints.* J Chem Inf Model, 2010. **50**(5): p. 742-54.

212. Han, L., Y. Wang, and S.H. Bryant, *Developing and validating predictive decision tree models from mining chemical structural fingerprints and high-throughput screening data in PubChem.* BMC Bioinformatics, 2008. **9**: p. 401.

213. Durant, J.L., et al., *Reoptimization of MDL keys for use in drug discovery.* J Chem Inf Comput Sci, 2002. **42**(6): p. 1273-80.

214. Duesbury, E., J. Holliday, and P. Willett, *Maximum common substructure-based data fusion in similarity searching.* J Chem Inf Model, 2015. **55**(2): p. 222-30.

215. Wang, Q., et al., *In silico prediction of serious eye irritation or corrosion potential of chemicals.* RSC advances, 2017. **7**(11): p. 6697-6703.

216. Kuhn, M., *Building predictive models in R using the caret package.* J Stat Softw, 2008. **28**(5): p. 1-26.

217. Rácz, A., D. Bajusz, and K. Héberger, *Modelling methods and cross-validation variants in QSAR: a multi-level analysis($).* SAR QSAR Environ Res, 2018. **29**(9): p. 661-674.

218. Chicco, D. and G. Jurman, *The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation.* BMC Genomics, 2020. **21**(1): p. 6.

219. Gevrey, M., I. Dimopoulos, and S. Lek, *Review and comparison of methods to study the contribution of variables in artificial neural network models.* Ecological modelling, 2003. **160**(3): p. 249-264.

220. Ferrari, T., et al., *Automatic knowledge extraction from chemical structures: the case of mutagenicity prediction.* SAR QSAR Environ Res, 2013. **24**(5): p. 365-83.

221. Keiser, M.J., et al., *Predicting new molecular targets for known drugs.* Nature, 2009. **462**(7270): p. 175-81.

222. Mohd Fauzi, F., et al., *Chemogenomics approaches to rationalizing the mode-of-action of traditional Chinese and Ayurvedic medicines.* J Chem Inf Model, 2013. **53**(3): p. 661-73.

223. Zhang, G.B., et al., *Actions of Huangqi decoction against rat liver fibrosis: a gene expression profiling analysis.* Chin Med, 2015. **10**: p. 39.

224. Li, W.K., et al., *Protective effect of herbal medicine Huangqi decoction against chronic cholestatic liver injury by inhibiting bile acid-stimulated inflammation in DDC-induced mice.* Phytomedicine, 2019. **62**: p. 152948.

225. Chen, E.Y., et al., *Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool.* BMC Bioinformatics, 2013. **14**: p. 128.

226. Clauset, A., M.E. Newman, and C. Moore, *Finding community structure in very large networks.* Phys Rev E Stat Nonlin Soft Matter Phys, 2004. **70**(6 Pt 2): p. 066111.

227. Chung, N.C., et al., *Jaccard/Tanimoto similarity test and estimation methods for biological presence-absence data.* BMC Bioinformatics, 2019. **20**(Suppl 15): p. 644.

228. Pan, X., et al., *Systematic review of the methodological quality of controlled trials evaluating Chinese herbal medicine in patients with rheumatoid arthritis.* BMJ open, 2017. **7**(3): p. e013242.

229. Yeh, H.Y., et al., *Predicting the Associations between Meridians and Chinese Traditional Medicine Using a Cost-Sensitive Graph Convolutional Neural Network.* Int J Environ Res Public Health, 2020. **17**(3).

230. Kubinyi, H., *Lipophilicity and drug activity.* Prog Drug Res, 1979. **23**: p. 97-198.

231. Zhou, S.S., et al., *Gut microbiota-involved mechanisms in enhancing systemic exposure of ginsenosides by coexisting polysaccharides in ginseng decoction.* Sci Rep, 2016. **6**: p. 22474.

232. Wu, X., et al., *Seeing the unseen of Chinese herbal medicine processing (Paozhi): advances in new perspectives.* Chin Med, 2018. **13**: p. 4.

233. Shu, B., Q. Shi, and Y.J. Wang, *Shen (Kidney)-tonifying principle for primary osteoporosis: to treat both the disease and the Chinese medicine syndrome.* Chin J Integr Med, 2015. **21**(9): p. 656-61.

234. Mateus, A., et al., *Thermal proteome profiling for interrogating protein interactions.* Mol Syst Biol, 2020. **16**(3): p. e9232.

235. Wang, Y., et al., *Mechanism of glycyrrhizin on ferroptosis during acute liver failure by inhibiting oxidative stress.* Mol Med Rep, 2019. **20**(5): p. 4081-4090.

236. Zhou, Y., et al., *Synergistic anti-liver fibrosis actions of total astragalus saponins and glycyrrhizic acid via TGF-β1/Smads signaling pathway modulation.* J Ethnopharmacol, 2016. **190**: p. 83-90.

237. Prasad, S., N. Kalra, and Y. Shukla, *Hepatoprotective effects of lupeol and mango pulp extract of carcinogen induced alteration in Swiss albino mice.* Mol Nutr Food Res, 2007. **51**(3): p. 352-9.