# IMPPAT: A curated database of Indian Medicinal Plants, Phytochemistry And Therapeutics

**Karthikeyan Mohanraj[1,#], Bagavathy Shanmugam Karthikeyan[1,#], R.P. Vivek-Ananth[1,#], R.P. Bharath Chand[1], S.R. Aparna[2], P. Mangalapandi[1], Areejit Samal[1,*]**

[1]The Institute of Mathematical Sciences (IMSc), Homi Bhabha National Institute, Chennai 600113, India

[2]Stella Maris College, Chennai 600086, India

[#]M.K., B.S.K. and R.P.V. contributed equally to this work

[*]Corresponding author: asamal@imsc.res.in

## Abstract

Phytochemical constituents of medicinal plants encompass a diverse space of chemical scaffolds which can be used for rational design of novel drugs. India is rich with a flora of indigenous medicinal plants that have been used for centuries in traditional Indian medicine to treat human maladies. A comprehensive online database on the phytochemistry of Indian medicinal plants will enable the application of systems biology and cheminformatic approaches towards natural product based drug discovery. In this direction, we here present, IMPPAT, a manually curated database of Indian Medicinal Plants, Phytochemistry, And Therapeutics. IMPPAT contains 1742 Indian medicinal plants, 9596 phytochemicals and 1124 therapeutic uses which span across 27074 plant-phytochemical associations and 11514 plant-therapeutic associations. Notably, the curation effort led to a non-redundant *in silico* chemical library of 9596 phytochemicals with standard chemical identifiers and structure information. Using cheminformatic approaches, we have computed the physicochemical properties and drug-likeliness of the phytochemicals in IMPPAT which led to a filtered subset of 960 potential druggable phytochemicals. Moreover, a comparative analysis against FDA approved drugs suggests that majority of the druggable phytochemicals in IMPPAT are good candidates for novel prospective drugs as they have little or no structural similarity with existing drugs. The IMPPAT database is openly accessible at: https://www.imsc.res.in/~asamal/resources/imppat/home.

## Introduction

Natural products continue to play a significant role in pharmaceutical industry[1-4] as new sources of drugs. However, recently there has been a decline in the number of marketable drugs derived from natural sources[3,4]. Furthermore, the majority of these drugs fall into already known structural scaffolds as due importance has not been given to unexplored sources of natural products for drug discovery[4]. As a result, lately, there has been significant interest in applying interdisciplinary approaches[5] such as text mining, natural language processing (NLP)[6], machine learning[7], cheminformatics[8], pharmacophore-based virtual screening[9,10], systems biology[11,12], systems pharmacology[13], network pharmacology[14] to expand the novel chemical scaffold libraries for drug discovery.

India is well known for its practice of traditional medicine and ethnopharmacology[15]. It is noteworthy that traditional Indian medicinal formulations are multi-component mixtures whose therapeutic use is based on empirical knowledge rather than a mechanistic understanding of the active ingredients in the mixture[15]. Until recently, knowledge of traditional Indian medicine including important medicinal plants and their formulations were buried within books such as Indian Materia Medica[16] and Ayurveda Materia Medica[17]. The nondigital nature of this information limited their effective use towards new drug discovery[5]. Hence, digitization of this knowledge into a comprehensive database on Indian medicinal plants, phytochemistry and ethnopharmacology will enable researchers to apply computational approaches towards drug discovery.

Availability of a curated database of plants, their associated natural products and a repository of their chemical structures, can help in *in silico* drug discovery. In this direction, there has been significant recent progress in the development of databases[18-25] on natural products with a focus on phytochemistry of edible and herbaceous plants. Examples of such databases include CVDHD[21], KNAPSACK[22], Nutrichem[18,19], Phytochemica[20], TCMID[23] and TCM-Mesh[24] which can facilitate virtual screening of prospective drug compounds or aid in the investigation of plant-disease associations. However, from the perspective of traditional Indian medicine, there have been relatively few efforts to build online databases that include Indian medicinal plants, their phytochemical constituents and therapeutic uses. Previously, Polur *et al*[26] compiled information on 295 ayurvedic Indian medicinal plants, their 1829 phytochemical constituents and therapeutic uses. Subsequently, Polur *et al*[26] studied the structural similarity between their library of 1829 phytochemicals and drugs in the DrugBank[27] database to predict biologically active natural compounds. Recently, the Phytochemica[20] database gathered information on 5 Indian medicinal plants and their 963 phytochemical constituents. In addition, Phytochemica[20] provided chemical structures and pharmacological properties of the phytochemicals within their database. Other efforts to build online databases for traditional Indian medicine has largely been limited to cataloguing medicinal plants and their therapeutic uses rather than capturing the phytochemical constituents that are vital for drug discovery. On the other hand, in contrast to the above mentioned online databases, more comprehensive databases are available for Chinese medicinal plants. For example, TCM-MeSH[24] is an online database for traditional Chinese medicine which captures phytochemical compositions and therapeutic uses for more than 6000 Chinese medicinal plants.

We therefore have built a manually curated database, IMPPAT, containing 1742 Indian Medicinal Plants, 9596 Phytochemical constituents, And 1124 Therapeutic uses. In addition, the IMPPAT database has linked Indian medicinal plants to 974 openly accessible traditional Indian medicinal formulations. Importantly, our curation efforts have led to a non-redundant *in silico* chemical library of 9596 phytochemical constituents for which we have computed physicochemical properties using cheminformatic tools[28]. We then employed cheminformatic approaches to evaluate the drug-likeliness of the phytochemicals in our *in silico* chemical library using multiple scoring schemes such as Lipinski's rule of five (RO5)[29], Oral PhysChem Score (Traffic Lights)[30], GlaxoSmithKline's (GSK's) 4/400[31], Pfizer's 3/75[32], Veber rule[33] and Egan rule[34]. We found a subset of 960 phytochemical constituents of Indian medicinal plants that are potentially druggable in our chemical library based on multiple scoring schemes. In summary, the IMPPAT database is a culmination of our efforts to digitize the wealth of information contained within traditional Indian medicine and provides an integrated platform where principles from systems biology and cheminformatics can be applied to accelerate natural product based drug discovery. IMPPAT is openly accessible at: https://www.imsc.res.in/~asamal/resources/imppat/home.

## Methods

### *Data collection, curation and processing*

**Curated list of Indian medicinal plants.** In the preliminary phase of the database construction (Figure 1), we compiled a comprehensive list of more than 5000 Indian medicinal plants based on information provided by the Foundation for Revitalisation of Local Health Traditions (FRLHT), Bengaluru (http://envis.frlht.org/), Central Institute of Medicinal and Aromatic Plants (CIMAP), Lucknow (http://cimap.res.in/) and Ministry of AYUSH, Government of India (http://ayush.gov.in/).

Due to the usage of multiple synonyms for medicinal plants across sources, the common names were converted into their scientific species names and the list was manually curated to remove redundancies. The Plant List database[35] (http://www.theplantlist.org/) was used for identifying synonyms of Indian medicinal plants.

**Phytochemical composition of Indian medicinal plants.** After compiling a comprehensive list of more than 5000 Indian medicinal plants, we mined literature to gather information on their phytochemical constituents (Figure 1). In the first stage of data mining, we focussed on specialized traditional Indian medicine books[36-45]. From these books[36-45], we gathered phytochemical composition for more than 1600 Indian medicinal plants. In the second stage, we gathered information from published databases of Indian medicinal plants. Phytochemica[20] is a dedicated electronic database for the phytochemical composition of Indian medicinal plants and contains information on physicochemical properties of 963 phytochemical constituents of 5 Indian medicinal plants. Another database described in Polur *et al*[26] had compiled information on 1829 phytochemical constituents of 295 ayurvedic Indian medicinal plants[26]. While this list is no longer publicly available, the Nutrichem[18,19] database on phytochemical composition and therapeutic uses of plant-based food products has incorporated the information compiled by Polur *et al*[26]. From the Phytochemica[20] and Nutrichem[18,19] databases, we gathered information on the phytochemical composition of more than 400 Indian medicinal plants. Note that our comprehensive list covers a wide spectrum of Indian medicinal plants which includes apart from Ayurveda, other systems of traditional Indian medicine such as Siddha and Unani. In the third stage of data mining for phytochemical composition, we performed text mining of abstracts from published research articles in PubMed[46] using natural language processing (NLP)[47]. Using in-house Python scripts, we identified keywords in PubMed abstracts which imply plant-phytochemical associations. We then used the selected keywords to mine PubMed abstracts to identify and incorporate additional references for plant-phytochemical associations in our database. In total, our database captures the phytochemical composition of 1742 Indian medicinal plants (Supplementary Table S1). The literature references for plant-phytochemical associations are listed in our database in the form of ISBN or DOI identifiers for books and PubMed identifiers (PMIDs) for journal articles.

**Annotation, curation and filtering of identified phytochemicals.** An overarching goal of this work is to create a platform for exploring the chemistry of the phytochemical constituents of Indian medicinal plants. Evaluation of the phytochemical constituents of Indian medicinal plants for their druggability or drug-likeliness will facilitate the identification of molecules for drug discovery. We would like to emphasize that synonymous chemical names are pervasive across the literature on traditional Indian medicine which were mined to construct this database. In order to remove redundancy, we manually annotated the common names of phytochemical constituents of Indian medicinal plants compiled from literature sources with documented synonyms and standard chemical identifiers (Figure 1) from Pubchem[48], CHEBI[49], CAS (https://www.cas.org/), CHEMSPIDER[50], KNAPSACK[51], CHEMFACES (http://www.chemfaces.com), FOODB (http://foodb.ca/), NIST Chemistry webbook[52] and Human Metabolome database (HMDB)[53]. While assigning standard identifiers to phytochemicals in our database, we have chosen the following priority order: Pubchem[48], CHEBI[49], CAS, CHEMSPIDER[50], KNAPSACK[51], CHEMFACES, FOODB, NIST Chemistry webbook[52] and HMDB[53]. We highlight that this extensive manual curation effort led to the mapping of more than 15000 common names of phytochemicals used across literature sources to a unique set of 9596 standard chemical identifiers. Phytochemicals which could not be mapped to standard chemical identifiers were excluded from our finalized database. Our choice to include only phytochemicals with standard identifiers and structure information was dictated by our goal to investigate the chemistry and druggability of phytochemical constituents of Indian medicinal plants. This largely manual effort to compile a non-redundant chemical library of 9596 phytochemical constituents of Indian medicinal plants with standard identifiers and structure information will serve as valuable resource for natural product based drug discovery in future. Moreover, the use of standard chemical identifiers will enable effortless integration of our IMPPAT database with other data sources.

**Therapeutic uses of Indian medicinal plants.** Another goal of our database is to compile ethnopharmacological information on Indian medicinal plants. Towards this goal, we manually compiled the medicinal (therapeutic) uses of Indian medicinal plants from books on Indian traditional medicine[36-45,52,54-70]. Apart from books, Polur *et al*[26] had previously compiled a list of therapeutic uses for 295 ayurvedic Indian medicinal plants, and this information was extracted from the Nutrichem[18,19] database. To ensure high quality, we manually curated information on therapeutic uses of Indian medicinal plants and consciously avoided automated text mining to retrieve additional information on plant-therapeutic associations. Furthermore, we manually annotated and standardized the compiled therapeutic uses of Indian medicinal plants from the above sources with identifiers from the Disease Ontology[71], Online Mendelian Inheritance in Man (OMIM)[72], Unified Medical Language System (UMLS)[73] and Medical Subject Headings (MeSH)[74] databases. To the best of our knowledge, this is the first large-scale attempt to link the ethnopharmacological information on Indian medicinal plants with standardized vocabulary in modern medicine.

**Traditional formulations of Indian medicinal plants.** Traditional knowledge digital library (TKDL) (http://www.tkdl.res.in) is a knowledgebase of traditional Indian medicinal formulations. According to TKDL, there are more than 250000 formulations of Ayurveda, Siddha and Unani of which 1200 representative formulations are openly accessible via their database. To exhibit the broader utility of our database to phytopharmacology, we have also compiled and curated the subset of 1200 openly accessible formulations in TKDL which contain at least one of the 1742 Indian medicinal plants in our database. This process led to associations between 321 Indian medicinal plants in our database and 974 traditional Indian medicinal formulations which are openly accessible through TKDL database (Figure 1). We emphasize that our database has only incorporated open digital information on traditional Indian medicinal formulations from TKDL database. However, we are aware of the vast literature[16,17,75] on traditional Indian medicinal formulations, especially in books, and in the future, a significant effort will be needed to digitize and integrate such information into our database.

*Database management and network visualization*

To construct this database, the compiled and curated data was integrated using MySQL (https://www.mysql.com/), a relational database management system which serves as a back-end for our resource. The web interface for the database was built using Drupal (https://www.drupal.org/), a PHP-based content management system hosted on Apache server with the MySQL database in the back-end. Users can browse or query our database using the scientific names of Indian medicinal plants, standard identifiers for phytochemicals, or associated therapeutic uses (Figure 2). Further we have integrated the Cytoscape.js application[76] (http://js.cytoscape.org/) into our web interface which enables visualization of plant-phytochemical associations and plant-therapeutic associations in the form of a network. The Cytoscape network visualization displays different types of nodes such as plant, phytochemical, therapeutic use and traditional medicinal formulations in different shapes and colours. Finally, the association network can be downloaded as a tab-separated list using the available export option in our database (Figure 2).

*Computation of physicochemical properties, druggability and similarity of phytochemicals*

**Physicochemical properties and druggability.** We used FAF-Drugs4 web-service[28] to compute the following physicochemical properties of the phytochemicals: molecular weight, partition coefficient, solubility in water, topological polar surface area, charge of the compound, number of hydrogen bond donors and acceptors, number of rotatable and rigid bonds, number of hetero- and heavy atoms, and number of stereocenters. FAF-Drugs4 web-service[28] tested the druggability of the phytochemicals based on multiple scoring schemes, namely, Lipinski's rule of five (RO5)[29], Oral PhysChem Score (Traffic Lights)[30], GlaxoSmithKline's (GSK's) 4/400[31], Pfizer's 3/75[32], Veber rule[33] and Egan rule[34]. We filtered phytochemicals with no RO5 violation, net Traffic Lights value of zero and satisfying GSK's 4/400, Pfizer's 3/75, Veber rule and Egan rule as *druggable*. We further computed the weighted quantitative estimate of drug-likeness (QEDw)[77] score using FAF-QED web-service[28] for the filtered list of druggable phytochemicals.

**Similarity of phytochemicals.** Tanimoto coefficient (Tc)[78] is a widely used measure to compute structural similarity between chemicals[79]. To evaluate the structural similarity of chemicals within our database to known drugs using Tc, we employed two molecular fingerprints: (a) Extended Circular Fingerprints (ECFP4)[80] applying Morgan algorithm[81] with radius value of 2 as implemented in RDKit[82], and (b) MACCS keys based fingerprint. We employed the open source package, RDKit[82], to compute molecular fingerprints and Tc between pairs of chemical structures. To identify structural similarity between chemicals, a stringent cut-off of Tc $\geq$ 0.5 was used while employing ECFP4 and a cut-off of Tc $\geq$ 0.85 was used while employing MACCS keys. Our selection of Tc cut-offs for ECFP4 and MACCS keys based computations was motivated by the recent work of Jasial *et al*[83].

We obtained a list of 2069 FDA approved drugs from DrugBank[27] and computed their structural similarity with our druggable phytochemicals using both ECFP4 and MACCS keys based molecular fingerprints. Note that ECFP4 molecular fingerprints were used to create the chemical similarity network of the druggable phytochemicals with QEDw score $\geq$ 0.9. Besides quantifying the structural similarity based on the Tc of phytochemicals, we have employed principal component analysis (PCA) to explore possible relationships between druggable phytochemicals with QEDw score $\geq$ 0.9 based on their physicochemical properties.

## Results

### *Web-interface of the database*

The IMPPAT database captures information on three types of associations for Indian medicinal plants: phytochemical composition, therapeutic uses, and traditional medicinal formulations (Figure 1). The web-interface of the database enables users to query for each of these associations using (a) scientific names of plants, (b) standard chemical identifiers of phytochemical constituents, (c) therapeutic uses, or (d) formulation identifiers (Figure 2). The web-interface displays the result of user queries for these associations in two ways: (a) A table of associations with references to literature sources, and (b) A network visualization of the associations which is powered by Cytoscape.js[76] (Figure 2). In addition, users can also download the result of their queries for different associations of medicinal plants as a tab-separated list using the available export option in the web interface. In the results page of queries for plant-phytochemical associations, users can click each phytochemical name or identifier to navigate to a separate page containing detailed information such as chemical structure, alternate chemical names or identifiers, computed physicochemical properties, computed druggability scores and the option to download the chemical structure file in SDF format (Figure 2; Methods). Queries for plant-therapeutic associations leads to a page where users can also obtain the disease ontology identifiers corresponding to therapeutic uses (Figure 2; Methods). In the results page of queries for plant-formulation associations, users can click the medicinal formulation identifiers to navigate to the corresponding page in the TKDL database. Moreover, in the query page of IMPPAT database, users can use advanced search options (Figure 2) to filter phytochemicals based on physicochemical properties (e.g., molecular weight, number of hydrogen bond acceptors) or satisfying various druggability scores (e.g. RO5, Traffic Lights).

*Network of plant-phytochemical associations, plant-therapeutic use associations, and plant-traditional medicinal formulation associations*

IMPPAT database contains information on the phytochemical composition and therapeutic uses of 1742 Indian medicinal plants (Supplementary Table S1). Of the 134 Indian medicinal plants in the priority list of Ministry of AYUSH, Government of India, 116 Indian medicinal plants are contained in our database (Supplementary Table S1). In addition, we identified 15 Indian medicinal plants in our database that appear in the red list of the International Union of Conservation of Nature (IUCN) as either near threatened, vulnerable, endangered or critically endangered (http://www.iucnredlist.org/). IMPPAT captures information on 27074 plant-phytochemical associations which encompasses 1742 Indian medicinal plants and their 9596 phytochemical constituents. Among the 1742 Indian medicinal plants in our database, *Catharanthus roseus* has the highest number of phytochemical associations. In Figure 3A, we show a histogram of the occurrence of phytochemicals across 1742 Indian medicinal plants in our database. From this figure, it is seen that the majority of phytochemicals are found in less than 5 Indian medicinal plants while only a handful of phytochemicals are found in more than 200 Indian medicinal plants.

IMPPAT also captures information on 11514 plant-therapeutic use associations which encompasses 1742 Indian medicinal plants and 1124 therapeutic uses. In Figure 3B, we show a histogram of the number of therapeutic uses per Indian medicinal plant in our database. From this figure, it is seen that the majority of Indian medicinal plants have less than 10 documented therapeutic uses while a small fraction of Indian medicinal plants have more than 20 therapeutic uses in our database. Among the 1742 Indian medicinal plants in our database, *Ginkgo biloba*, *Panax ginseng* and *Allium sativum* have the largest number of documented therapeutic uses. Lastly, IMPPAT also captures information on 5069 plant-formulation associations which encompasses 321 Indian medicinal plants in our database and 974 traditional Indian medicinal formulations which are openly accessible from the TKDL database (Methods).

*Druggability analysis of phytochemical constituents of Indian medicinal plants*

**Druggable phytochemicals.** We evaluated the druggability of 9596 phytochemicals in IMPPAT database based on multiple rules or scoring schemes, namely, RO5[29], Traffic Lights[30], GSK's 4/400[31], Pfizer's 3/75[32], Veber rule[33] and Egan rule[34] which were computed using FAF-Drugs4 web-service[28] (Methods). The horizontal bar plot in Figure 4A gives the number of phytochemicals in IMPPAT that satisfy different druggability scores. From this figure, it is seen that the majority of our phytochemicals satisfy Veber or Egan rules in comparison to Pfizer's 3/75 rule or net Traffic Lights value of zero. Furthermore, we find that the same set of 8712 phytochemicals in IMPPAT satisfy both the Veber rule and Egan rule for drug-likeliness. The vertical bar plot of Figure 4A shows the overlap between sets of phytochemicals that satisfy different druggability scores. We found that 960 out of 9596 phytochemicals in IMPPAT database satisfy all evaluated druggability scores (Figure 4A). Subsequently, we designated this filtered list of 960 phytochemicals as *druggable*. Among the 1742 Indian medicinal plants in our database, *Brassica oleracea, Catharanthus roseus, Zea mays, Oryza sativa, Vigna radiate, Pisum sativum, Anethum sowa, Allium cepa, Cassia obtusifolia* and *Camellia sinensis* produce the highest number of druggable phytochemicals (Table 1). In Figure 4B, we show the classification of the 960 druggable phytochemicals into broad classes similar to the classification of natural products in NPACT[25] database. It is seen that the subset of 960 druggable phytochemicals is enriched in flavonoids and terpenoids. In Figure 4C, we show the distribution of weighted quantitative estimation of drug-likeness (QEDw) score[77] for the 960 druggable phytochemicals. From this figure, it is seen that 14 druggable phytochemical have a QEDw score $\geq 0.9$ and 98 druggable phytochemicals have a QEDw score $\geq 0.8$.

**Overlap with approved drugs space.** We obtained the structures of 2069 FDA approved drugs from the DrugBank[27] database. By investigating the structural similarity between FDA approved drugs and 960 druggable phytochemicals in our IMPPAT database, we found that 249 and 302 druggable phytochemicals are similar to FDA approved drugs based on ECFP4 or MACCS keys molecular fingerprints respectively (Figure 4D; Methods). Combined, ECFP4 and MACCS keys based fingerprints identified 369 out of 960 druggable phytochemicals that are similar to FDA approved drugs (Methods). Thus, almost 40% of the druggable phytochemicals in IMPPAT database are similar to at least one FDA approved drug which testifies to our systemic approach to identify potential druggable phytochemical constituents of Indian medicinal plants. Importantly, the remaining 591 druggable phytochemicals which have no similarity with any of the FDA approved drugs are novel candidates for designing new drugs based on natural products from Indian medicinal plants.

**Chemical similarity network of the most-druggable phytochemicals.** For subsequent analysis, we selected 14 druggable phytochemicals with QEDw score[77] $\geq 0.9$ which were designated as the *most-druggable* phytochemicals. Of these 14 phytochemicals, 12 were found to be similar to at least one of the FDA approved drugs based on either ECFP4 or MACCS keys based molecular fingerprint. The remaining 2 most-druggable phytochemicals, Onosmone (CID:102212116) and Truxillic acid (CID:78213), were found to have no similarity with any of the FDA approved drugs. In order to probe the structural diversity of these 14 most-druggable phytochemicals, we computed the Tc based on ECFP4 molecular fingerprint between all pairs of phytochemicals (Methods). In Figure 5A, we display the similarity matrix based on Tc for the 14 most-druggable phytochemicals. From this figure, it is seen that the majority of the Tc values are small in the similarity matrix implying high structural diversity. Moreover, the similarity matrix can be transformed into a similarity network of phytochemicals by using a stringent threshold value of Tc $\geq 0.5$ to determine edges in the graph (Figure 5B). We find that only 16 of the 91 possible edges between the 14 most-druggable phytochemicals are realized in the similarity network (Figure 5B). Furthermore, the similarity network can be partitioned into a large connected component (cluster) of 7 phytochemicals, a smaller connected component of 2 phytochemicals and 5 remaining isolated phytochemicals. We highlight that the 2 phytochemicals, Onosmone and Truxillic acid, that have no similarity with any of the FDA approved drugs are among the isolated nodes in the similarity network (Figure 5B). Based on plant-phytochemical associations in our database, Onosmone and Truxillic acid are phytochemical constituents of Indian medicinal plants, *Onosma echioides* and

*Erythroxylum coca*, respectively, and a survey of the literature shows that these phytochemicals are under active investigation for their therapeutic uses[84-88]. We also highlight that none of the 14 most-druggable phytochemicals are captured by Phytochemica[20] database while 6 of the 14 phytochemicals are captured by Nutrichem[18,19] database.

**Principal component analysis of the most-druggable phytochemicals based on their physicochemical properties.** In the last section, we studied the similarity between chemical structures of 14 most-druggable phytochemicals to find a large cluster of 7 phytochemicals with highly similar chemical structures. But it is well known that high similarity between chemical structures does not necessarily imply high similarity between chemical activities[89]. Thus, we here investigate the physicochemical properties of the 14 most-druggable phytochemicals. Note that we have used FAF-Drugs4 web-service[28] to compute several physicochemical properties including molecular weight, partition coefficient, solubility in water, topological polar surface area, charge of the compound, number of hydrogen bond donors and acceptors, number of rotatable and rigid bonds, number of hetero- and heavy atoms, and number of stereocenters for each phytochemical in IMPPAT database (Methods). In Figure 6, we show the distribution of four physicochemical properties, namely, molecular weight, number of hydrogen bond donors, number of hydrogen bond acceptors and topological polar surface area, across the 9596 phytochemicals in our database. We performed principal component analysis (PCA) of the 14 most-druggable phytochemicals based on their physiochemical properties (Figure 5C). In Figure 5C, the first and second principal components together explained more than 71% of the total variance in the dataset. We find that the 7 most-druggable phytochemicals which are clustered together in the structural similarity space (Figure 5B) are not clustered together in the physicochemical or chemical activity space (Figure 5C). These observations based on limited analysis of 14 most-druggable phytochemicals suggest that a combined exploration of chemical similarity space and physicochemical or chemical activity space of phytochemical constituents of Indian medicinal plants will be required in the future to identify and design novel drugs.

## Discussion and future directions

In the 21$^{st}$ century, there is immense interest within academia and pharmaceutical industry to incorporate systems biology approaches to accelerate the drug discovery pipeline which has led to the emergence of sub-disciplines such as systems pharmacology[13] and network pharmacology[14]. Likewise cheminformatics can accelerate drug discovery by aiding in the rational design of robust chemical scaffolds from diverse natural sources[5]. Recently, Lagunin *et al*[5] have reviewed more than 50 existing bioinformatics resources for natural product based drug discovery. In their review, Lagunin *et al*[5] noted that most existing databases on medicinal plants and associated phytochemistry do not provide a platform to integrate systems biology and cheminformatics approaches which impedes natural products based drug discovery. Towards this goal, we here incorporate principles from systems biology and cheminformatics to build an extensive resource on phytochemistry and ethnopharmacology of Indian medicinal plants. Here we present, IMPPAT, a curated database of 1742 Indian Medicinal Plants, 9596 Phytochemical constituents, And 1124 Therapeutic uses. Importantly, IMPPAT provides a unifying platform for the application of system-level approaches to elucidate mechanistic links between phytochemical constituents of Indian medicinal plants and their therapeutic action.

Using cheminformatic approaches, we found that 960 of the 9596 phytochemical constituents of Indian medicinal plants in our database are potentially druggable based on multiple scoring schemes. Of the 960 phytochemicals which satisfy all druggability scores evaluated here, a subset of 14 phytochemicals were found to have a QEDw score[77] $\geq 0.9$ (Figure 5). Interestingly, the occurrence of these 14 most-druggable phytochemicals across 1742 Indian medicinal plants in our database is very rare with none of the 14 phytochemicals being found in more than 3 Indian medicinal plants. Specifically, the 14 most-druggable phytochemicals are constituents of only 17 Indian medicinal plants in our database. Also, 4 of the 14 most-druggable phytochemicals are constituents of 3 phylogenetically close Indian medicinal plants, *Iris germanica*, *Iris nepalensis* and *Iris kemaonensis*, which are from the same genus. However, we find that only 2 out of 17 Indian medicinal plants that produce the 14 most-druggable phytochemicals are in the priority list of Ministry of AYUSH, Government of India. This analysis suggests a possible revision in the AYUSH priority list to include the remaining 15 Indian medicinal plants that produced the majority of the most-druggable phytochemicals in our database. Thus, our resource will facilitate rational design of scaffolds for new drugs based on natural products and future expansion of the chemotaxonomy[90] of Indian medicinal plants.

In the future, we hope to update IMPPAT database with the following additional information. Firstly, it will be important to link the phytochemical constituents of the Indian medicinal plants with their gene or protein targets. Such target information is vital to obtain a mechanistic understanding of either the therapeutic action or toxic effects of Indian medicinal plants. For example, TCM-Mesh[24], a traditional Chinese medicine database has gathered gene or protein target information for phytochemical constituents of Chinese medicinal plants using a network pharmacology approach. Importantly, information on gene or protein targets of phytochemicals will also enable pathway level assessment of the therapeutic action of medicinal plants which will help design robust drug scaffolds for many complex diseases. Secondly, it will be important to update our database with more detailed information on the parts of the Indian medicinal plants such as leaves, stem or root, that produce the different phytochemical constituents. Such detailed information on the phytochemical composition of parts of Indian medicinal plants will be crucial for evaluating and developing traditional Indian medicine formulations[75]. However, significant manual curation and literature mining will be needed to expand our database to include the phytochemical composition of the different parts of 1742 Indian medicinal plants which is beyond the scope of the present work. Thirdly, it will be important to enrich our database by incorporating more traditional Indian medicinal formulations. For example, TKDL (http://www.tkdl.res.in) has made only 1200 of their documented 250000 traditional Indian medicinal formulations openly accessible, and future efforts to associate this wealth of information to our database will shed mechanistic information on the therapeutic action of traditional formulations. Lastly, it will be important to perform a comparative analysis of the phytochemical composition of Indian medicinal plants with those of Chinese medicinal plants. Such a comparative analysis will shed light on phytochemicals and druggable scaffolds exclusive to Indian medicinal plants.

## Acknowledgments

We would like to thank Gopal C. Nanda and the staff of Achanta Lakshmipathi Research Centre for Ayurveda, Chennai for discussions and help in accessing relevant literature. We also thank S. Krishnaswamy and James Craig for discussions and the library staff of The Institute of Mathematical Sciences, Chennai for facilitating access to scientific literature. AS acknowledges financial support from Department of Science and Technology (DST) India through the award of a start-up grant (YSS/2015/000060) and Ramanujan fellowship (SB/S2/RJN-006/2014), Max Planck Society Germany through the award of a Max Planck Partner Group, and Department of Atomic Energy (DAE) India. The funders have no role in study design, data collection, data analysis, manuscript preparation or decision to publish.

## Author contributions

A.S., B.S.K., R.P.V. and M.K. designed research. M.K., B.S.K., R.P.V, R.P.B., S.R.A. and P.M. compiled and curated data from various sources. M.K. designed the database platform and visual interface. R.P.V. performed the cheminformatic analysis. B.S.K., R.P.V. and A.S. wrote the manuscript. All authors have read and approved the manuscript.

## Competing Interests

The authors declare that they have no competing interests.

## Figure captions

**Figure 1: Schematic overview of the IMPPAT database construction pipeline.** Briefly, we first compiled a comprehensive list of Indian medicinal plants from various sources. We next mined specialized books on Indian traditional medicine, existing databases and PubMed abstracts of journal articles to gather information on phytochemical constituents of Indian medicinal plants. We then manually annotated, curated and indexed names of identified phytochemicals with standard identifiers to build a non-redundant library of phytochemicals. This manual curation effort led to a unique list of plant-phytochemical associations. Subsequently, we gathered ethnopharmacological information from books on traditional Indian medicine to build a unique list of plant-therapeutic use associations. We also extracted publicly accessible information on traditional medicine formulations from TKDL database to build a list of plant-formulation associations. Lastly, we have used cheminformatic tools to compute the physicochemical properties and drug-likeliness of phytochemical constituents.

**Figure 2: Web-interface of the IMPPAT database.** (A) Snapshot of the result of a standard query for phytochemical constituents of an Indian medicinal plant. In this example, we show the plant-phytochemical association for *Ocimum tenuiflorum*, commonly known as Tulsi, from IMPPAT database. (B) Snapshot of the dedicated page containing detailed information on chemical structure, physicochemical properties and druggability scores for a chosen phytochemical. From the dedicated page for each phytochemical, users can download the structure of the phytochemical in the form of a SDF or MOL file. (C) Snapshot of the result of a standard query for therapeutic uses of an Indian medicinal plant. In this example, we show the therapeutic uses of *Zingiber officinale*, commonly known as Ginger, from IMPPAT database. (D) Snapshot of the advanced search options which enable users to filter phytochemicals based on their physiochemical properties or druggability scores.

**Figure 3: Occurrence of phytochemicals and therapeutic uses of Indian medicinal plants in IMPPAT database.** (A) Histogram of the occurrence of 9596 phytochemicals across 1742 Indian medicinal plants in our database. (B) Histogram of the number of therapeutic uses per Indian medicinal plant in our database.

**Figure 4: Druggability analysis of phytochemicals in IMPPAT database.** (A) Evaluation of drug-likeliness of phytochemicals based on multiple scores. The horizontal bar plot shows the number of phytochemicals in the IMPPAT database that satisfy different druggability scores (Methods). The vertical bar plot shows the overlap between sets of phytochemicals that satisfy different druggability scores. The pink bar in the vertical plot gives the 960 phytochemicals which satisfy all druggability scores. This plot was generated using UpSetR[91] package. (B) Classification of the 960 druggable phytochemicals into broad chemical classes. (C) Distribution of weighted quantitative estimate of drug-likeness (QEDw)[77] score for the 960 phytochemicals which satisfy all druggability scores. (D) Venn diagrams summarizing structural similarity analysis of 960 druggable phytochemicals in IMPPAT database and FDA approved drugs. Based on ECFP4 and MACCS keys molecular fingerprints, 249 and 302 druggable phytochemicals, respectively, were found to be similar to FDA approved drugs.

**Figure 5: Structural similarity and physicochemical properties of most-druggable phytochemicals in IMPPAT database.** (A) Similarity matrix for the 14 most-druggable phytochemicals with QEDw score $\geq 0.9$ based on Tanimoto coefficient (Tc) between pairs of chemicals computed using ECFP4 molecular fingerprints. (B) Similarity network for the 14 most-druggable phytochemicals constructed using a stringent threshold value of Tc $\geq 0.5$ to determine edges in the graph. We find that the similarity network can be partitioned into a large connected component of 7 phytochemicals, a smaller connected component of 2 phytochemicals and 5 isolated phytochemicals. (C) Principal component analysis (PCA) of the 14 most-druggable phytochemicals based on their physicochemical properties. The first and second principal components can together explain more than 71% of the total variance in the dataset.

**Figure 6: Distribution of physicochemical properties across the 9596 phytochemicals in IMPPAT database.** (A) Molecular weight. (B) Number of hydrogen bond donors. (C) Number of hydrogen bond acceptors. (D) Topological polar surface area.

## Tables

**Table 1:** List of 10 Indian medicinal plants with the highest number of druggable phytochemicals in the IMPPAT database. The table also lists the number of phytochemical constituents of the Indian medicinal plants.

| **Indian medicinal plant** | **Number of druggable phytochemicals** | **Number of phytochemicals** |
|---|---|---|
| *Brassica oleracea* | 43 | 257 |
| *Catharanthus roseus* | 42 | 403 |
| *Zea mays* | 29 | 116 |
| *Oryza sativa* | 23 | 181 |
| *Vigna radiata* | 23 | 139 |
| *Pisum sativum* | 22 | 112 |
| *Anethum sowa* | 21 | 104 |
| *Allium cepa* | 20 | 89 |
| *Cassia obtusifolia* | 20 | 70 |
| *Camellia sinensis* | 19 | 101 |

## Supplementary Material

**Supplementary Table S1:** List of 1742 Indian medicinal plants with information on the phytochemical composition and therapeutic uses in IMPPAT database. Of the 1742 Indian medicinal plants in IMPPAT database, 116 are on the priority list of Ministry of AYUSH, Government of India and 15 are on the red list of the International Union of Conservation of Nature (IUCN).

## References

1. Koehn, F. E. & Carter, G. T. The evolving role of natural products in drug discovery. *Nat Rev Drug Discov* **4**, 206-220 (2005).

2. Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *J Nat Prod* **79**, 629-661 (2016).

3. Li, J. W. & Vederas, J. C. Drug discovery and natural products: end of an era or an endless frontier? *Science* **325**, 161-165 (2009).

4. Pye, C. R., Bertin, M. J., Lokey, R. S., Gerwick, W. H. & Linington, R. G. Retrospective analysis of natural products provides insights for future discovery trends. *Proc Natl Acad Sci U S A* **114**, 5601-5606 (2017).

5. Lagunin, A. A. *et al.* Chemo- and bioinformatics resources for in silico drug discovery from medicinal plants beyond their traditional use: a critical review. *Nat Prod Rep* **31**, 1585-1611 (2014).

6. Rindflesch, T. C., Tanabe, L., Weinstein, J. N. & Hunter, L. EDGAR: extraction of drugs, genes and relations from the biomedical literature. *Pac Symp Biocomput*, 517-528 (2000).

7. Burbidge, R., Trotter, M., Buxton, B. & Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput Chem* **26**, 5-14 (2001).

8. Olsson, T. & Oprea, T. I. Cheminformatics: a tool for decision-makers in drug discovery. *Curr Opin Drug Discov Devel* **4**, 308-313 (2001).

9. Isgut, M. *et al.* Application of Combination High-Throughput Phenotypic Screening and Target Identification Methods for the Discovery of Natural Product-Based Combination Drugs. *Med Res Rev* (2017).

10. Steindl, T. M., Schuster, D., Laggner, C. & Langer, T. Parallel screening: a novel concept in pharmacophore modeling and virtual screening. *J Chem Inf Model* **46**, 2146-2157 (2006).

11. Butcher, E. C. Can cell systems biology rescue drug discovery? *Nat Rev Drug Discov* **4**, 461-467 (2005).

12. Pujol, A., Mosca, R., Farres, J. & Aloy, P. Unveiling the role of network and systems biology in drug discovery. *Trends Pharmacol Sci* **31**, 115-123 (2010).

13. Xie, L., Draizen, E. J. & Bourne, P. E. Harnessing Big Data for Systems Pharmacology. *Annual Review of Pharmacology and Toxicology* **57**, 245-262 (2017).

14. Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol* **4**, 682-690 (2008).

15. Pandey, M. M., Rastogi, S. & Rawat, A. K. Indian traditional ayurvedic system of medicine and nutritional supplementation. *Evid Based Complement Alternat Med* **2013**, 376327 (2013).

16. Nadkarni, M. K. & Nadkarni, A. K. *Indian materia medica*. (Popular Book Depot, 1955).

17. Dash, V. B. & Kashyap, V. L. *Ayurveda Materia Medica*. 711 (Concept Publishing Company, 1999).

18. Jensen, K., Panagiotou, G. & Kouskoumvekaki, I. Integrated text mining and chemoinformatics analysis associates diet to health benefit at molecular level. *PLoS Comput Biol* **10**, e1003432 (2014).

19. Jensen, K., Panagiotou, G. & Kouskoumvekaki, I. NutriChem: a systems chemical biology resource to explore the medicinal value of plant-based foods. *Nucleic Acids Res* **43**, D940-945 (2015).

20. Pathania, S., Ramakrishnan, S. M. & Bagler, G. Phytochemica: a platform to explore phytochemicals of medicinal plants. *Database (Oxford)* **2015** (2015).

21. Gu, J., Gui, Y., Chen, L., Yuan, G. & Xu, X. CVDHD: a cardiovascular disease herbal database for drug discovery and network pharmacology. *J Cheminform* **5**, 51 (2013).

22. Afendi, F. M. *et al.* KNApSAcK family databases: integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol* **53**, e1 (2012).

23. Xue, R. *et al.* TCMID: Traditional Chinese Medicine integrative database for herb molecular mechanism analysis. *Nucleic Acids Res* **41**, D1089-1095 (2013).

24. Zhang, R. Z., Yu, S. J., Bai, H. & Ning, K. TCM-Mesh: The database and analytical system for network pharmacology analysis for TCM preparations. *Sci Rep* **7**, 2821 (2017).

25. Mangal, M., Sagar, P., Singh, H., Raghava, G. P. & Agarwal, S. M. NPACT: Naturally Occurring Plant-based Anti-cancer Compound-Activity-Target database. *Nucleic Acids Res* **41**, D1124-1129 (2013).

26. Polur, H., Joshi, T., Workman, C. T., Lavekar, G. & Kouskoumvekaki, I. Back to the Roots: Prediction of Biologically Active Natural Products from Ayurveda Traditional Medicine. *Mol Inform* **30**, 181-187 (2011).

27. Law, V. *et al.* DrugBank 4.0: shedding new light on drug metabolism. *Nucleic Acids Res* **42**, D1091-1097 (2014).

28. Lagorce, D., Bouslama, L., Becot, J., Miteva, M. A. & Villoutreix, B. O. FAF-Drugs4: free ADME-tox filtering computations for chemical biology and early stages drug discovery. *Bioinformatics* (2017).

29. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings1PII of original article: S0169-409X(96)00423-1. The article was originally published in Advanced Drug Delivery Reviews 23 (1997) 3–25.1. *Advanced Drug Delivery Reviews* **46**, 3-26 (2001).

30. Lobell, M. *et al.* In Silico ADMET Traffic Lights as a Tool for the Prioritization of HTS Hits. *ChemMedChem* **1**, 1229-1236 (2006).

31. Gleeson, M. P. Generation of a Set of Simple, Interpretable ADMET Rules of Thumb. *Journal of Medicinal Chemistry* **51**, 817-834 (2008).

32. Hughes, J. D. *et al.* Physiochemical drug properties associated with in vivo toxicological outcomes. *Bioorganic & Medicinal Chemistry Letters* **18**, 4872-4875 (2008).

33. Veber, D. F. *et al.* Molecular properties that influence the oral bioavailability of drug candidates. *Journal of medicinal chemistry* **45**, 2615-2623 (2002).

34. Egan, W. J., Merz, K. M. & Baldwin, J. J. Prediction of Drug Absorption Using Multivariate Statistics. *Journal of Medicinal Chemistry* **43**, 3867-3877 (2000).

35. Kalwij, J. M. Review of 'The Plant List, a working list of all plant species'. *Journal of Vegetation Science* **23**, 998-1002 (2012).

36. Khare, C. P. *Indian medicinal plants : an illustrated dictionary*. (Springer, 2007).

37. Sharma, R. & Gupta, T. *Encyclopaedia of Medicinal Plants*. Vol. I (Campus Books International, 2009).

38. Sharma, R. & Gupta, T. *Encyclopaedia of Medicinal Plants*. Vol. II (Campus Books International, 2009).

39. Sharma, R. & Gupta, T. *Encyclopaedia of Medicinal Plants*. Vol. III (Campus Books International, 2009).

40. Sharma, R. & Gupta, T. *Encyclopaedia of Medicinal Plants*. Vol. IV (Campus Books International, 2009).

41. Sharma, R. & Gupta, T. *Encyclopaedia of Medicinal Plants*. Vol. V (Campus Books International, 2009).

42. Sharma, R. & Gupta, T. *Encyclopaedia of Medicinal Plants*. Vol. VI (Campus Books International, 2009).

43. Sharma, R. & Gupta, T. *Encyclopaedia of Medicinal Plants*. Vol. VII (Campus Books International, 2009).

44. Sharma, R. & Gupta, T. *Encyclopaedia of Medicinal Plants*. Vol. VIII (Campus Books International, 2009).

45. Duke, J. A. *Handbook of phytochemical constituents of GRAS herbs and other economic plants*. (CRC Press, 1992).

46. Coordinators, N. R. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Res* **45**, D12-D17 (2017).

47. Bird, S., Klein, E. & Loper, E. *Natural Language Processing with Python*. (O'Reilly Media, Inc., 2009).

48. Kim, S. *et al.* PubChem Substance and Compound databases. *Nucleic Acids Res* **44**, D1202-1213 (2016).

49. Hastings, J. *et al.* The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res* **41**, D456-463 (2013).

50. Editorial: ChemSpider--a tool for Natural Products research. *Nat Prod Rep* **32**, 1163-1164 (2015).

51. Nakamura, K. *et al.* KNApSAcK-3D: a three-dimensional structure database of plant metabolites. *Plant Cell Physiol* **54**, e4 (2013).

52. P.J. Linstrom and W.G. Mallard, E. *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*. (National Institute of Standards and Technology, Gaithersburg MD, 20899).

53. Wishart, D. S. *et al.* HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res* **41**, D801-807 (2013).

54. Deorani, S. C. & Sharma, G. D. *Medicinal plants of Nagaland*. (Bishen Singh Mahendra Pal Singh, 2007).

55. Dhiman, A. K. & Hora, S. L. *Wild Medicinal Plants of India*. (BSMPS, 2005).

56. Govil, J. N. & Singh, V. K. *Recent progress in medicinal plants-Ethnomedicine and pharmacognosy*. Vol. I (Studium Press LLC, U.S.A., 2002).

57. Govil, J. N. & Singh, V. K. *Recent progress in medicinal plants-Ethnomedicine and pharmacognosy II*. Vol. VII (Studium Press LLC, U.S.A., 2003).

58. Govil, J. N. & Singh, V. K. *Recent progress in medicinal plants-Ethnomedicine and pharmacognosy IV*. Vol. XXII (Studium Press LLC, U.S.A., 2008).

59. Govil, J. N. & Singh, V. K. *Recent progress in medicinal plants- Phytopharmacology and Therapeurtic values II*. (Studium Press LLC, U.S.A., 2008).

60. Gupta, A. K., Tandon, N. & Sharma, M. *Quality Standards of Indian Medicinal Plants*. (Indian Council of Medical Research, 2006).

61. Kaushik, P. & Dhiman, A. K. *Medicinal plants and raw drugs of India*. (Bishen Singh Mahendra Pal Singh, 2000).

62. Kirtikar, K. R. & Basu, B. D. *Indian Medicinal Plants*. Vol. I ( Periodical Experts Book Agency, 2006).

63. Kirtikar, K. R. & Basu, B. D. *Indian Medicinal Plants*. Vol. II ( Periodical Experts Book Agency, 2006).

64. Kirtikar, K. R. & Basu, B. D. *Indian Medicinal Plants*. Vol. III ( Periodical Experts Book Agency, 2012).

65. Kshirsagar, R. D. & Singh, N. P. *Ethnobotany of Mysore and Coorg, Karnataka State*. (Bishen Singh Mahendra Pal Singh, 2007).

66. P., K. C. *Medicinal Plants of Indian Trans Himalaya: Focus on Tibetan use of Medicinal Resources*. (Bishen Singh Mahendra Pal Singh, 2003).

67. Pande, P. C., Tiwari, L. & Pande, H. C. *Folk-medicine and aromatic plants of Uttaranchal*. (Bishen Singh Mahendra Pal Singh, 2006).

68. Sharma, U. K. *Medicinal plants of Assam*. (2004).

69. Singh, K. K. & Kushal, K. *Ethnobotanical wisdom of Gaddi tribe in Western Himalaya*. (Bishen Singh Mahendra Pal Singh, 2000).

70. Viswanathan, M. B., Prem Kumar, E. H. & Ramesh, N. *Ethnobotany of the Kanis : Kalakkad-Mundanthurai Tiger Reserve in Tirunelveli district, Tamilnadu, India*. (Bishen Singh Mahendra Pal Singh, 2006).

71. Schriml, L. M. *et al.* Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Res* **40**, D940-946 (2012).

72. Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. & McKusick, V. A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res* **33**, D514-517 (2005).

73. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* **32**, D267-270 (2004).

74. Rogers, F. B. Medical subject headings. *Bull Med Libr Assoc* **51**, 114-116 (1963).

75. Dev, S. Ancient-modern concordance in Ayurvedic plants: some examples. *Environ Health Perspect* **107**, 783-789 (1999).

76. Franz, M. *et al.* Cytoscape.js: a graph theory library for visualisation and analysis. *Bioinformatics* **32**, 309-311 (2016).

77. Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature chemistry* **4**, 90-98 (2012).

78. Tanimoto, T. T. IBM Internal Report 17th Nov. (1957).

79. Bajusz, D., Racz, A. & Heberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J Cheminform* **7**, 20 (2015).

80. Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **50**, 742-754 (2010).

81. Morgan, H. L. The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *Journal of Chemical Documentation* **5**, 107-113 (1965).

82. RDKit: Open-source cheminformatics. <http://www.rdkit.org>.

83. Jasial, S., Hu, Y., Vogt, M. & Bajorath, J. Activity-relevant similarity values for fingerprints and implications for similarity searching. *F1000Research* **5**, Chem Inf Sci-591 (2016).

84. Ahmad, V. U. *et al.* A New Ketone and a Known Anticancer Triterpenoid from the Leaves of Onosma limitaneum. *Helvetica Chimica Acta* **88**, 309-311 (2005).

85. Kumar, N., Kumar, R. & Kishore, K. Onosma L.: A review of phytochemistry and ethnopharmacology. *Pharmacogn Rev* **7**, 140-151 (2013).

86. Marson, C. M. New and unusual scaffolds in medicinal chemistry. *Chem Soc Rev* **40**, 5514-5533 (2011).

87. Rupp, M. *et al.* From machine learning to natural product derivatives that selectively activate transcription factor PPARgamma. *ChemMedChem* **5**, 191-194 (2010).

88. Sokolova, A. P., Alla; Ardashov, K.; Shernyukov, A.; Gatilov, Y.; Yarovaya, O.; Tolstikova, O.; Salakhutdinov, N. Synthesis and analgesic activity of new α-truxillic acid derivatives with monoterpenoid fragments. *Medicinal Chemistry Research* **25**, 1608-1615 (2016).

89. Maggiora, G. M. On outliers and activity cliffs--why QSAR often disappoints. *J Chem Inf Model* **46**, 1535 (2006).

90. Verpoorte, R. Exploration of nature's chemodiversity: the role of secondary metabolites as leads in drug development. *Drug Discovery Today* **3**, 232-238 (1998).

91. Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938-2940 (2017).

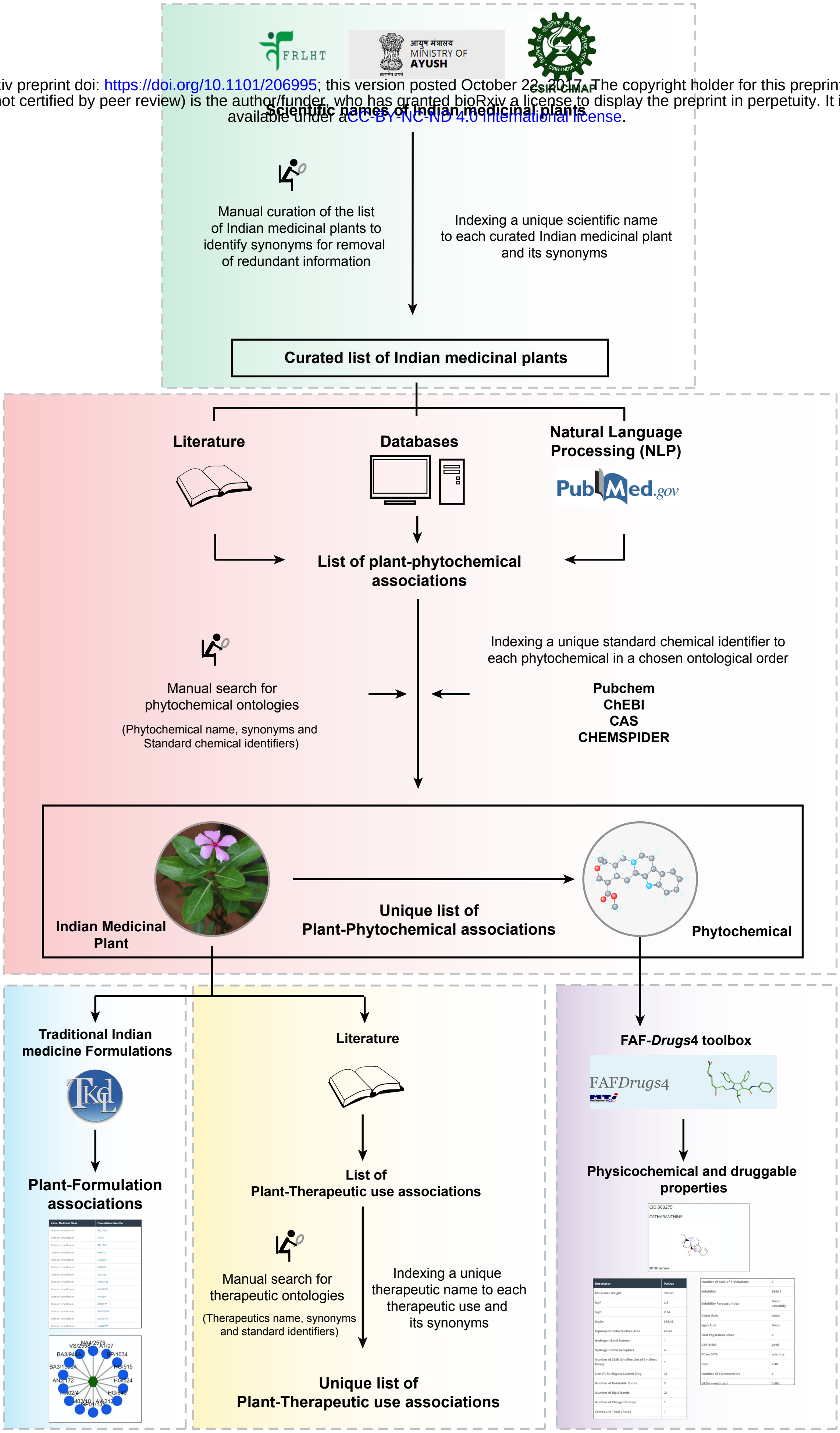
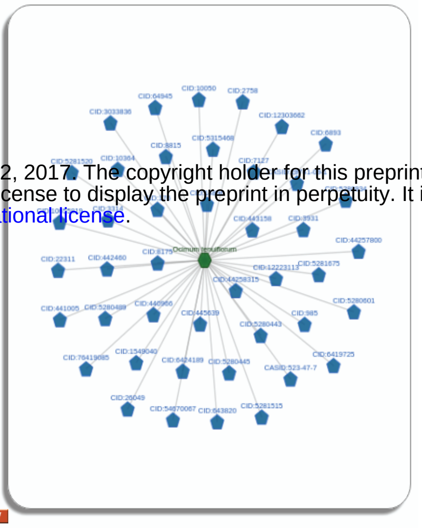

FRLHT
आयुष मंत्रालय
MINISTRY OF
AYUSH
CSIR-CIMAP
Scientific names of Indian medicinal plants
Manual curation of the list of Indian medicinal plants to identify synonyms for removal of redundant information
Indexing a unique scientific name to each curated Indian medicinal plant and its synonyms
Curated list of Indian medicinal plants
Literature
Databases
Natural Language Processing (NLP)
PubMed.gov
List of plant-phytochemical associations
Manual search for phytochemical ontologies
(Phytochemical name, synonyms and Standard chemical identifiers)
Indexing a unique standard chemical identifier to each phytochemical in a chosen ontological order
Pubchem
ChEBI
CAS
CHEMSPIDER
Indian Medicinal Plant
Unique list of Plant-Phytochemical associations
Phytochemical
Traditional Indian medicine Formulations
Plant-Formulation associations
Literature
List of Plant-Therapeutic use associations
Manual search for therapeutic ontologies
(Therapeutics name, synonyms and standard identifiers)
Indexing a unique therapeutic name to each therapeutic use and its synonyms
Unique list of Plant-Therapeutic use associations
FAF-Drugs4 toolbox
FAFDrugs4
Physicochemical and druggable properties

Figure 1

A

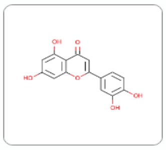
| INDIAN MEDICINAL PLANT | PHYTOCHEMICAL IDENTIFIER | PHYTOCHEMICAL NAME | REFERENCES |
|---|---|---|---|
| Ocimum tenuiflorum | CASID:29741-09-1 | Apigenin 7-glucuronide | ISBN:978-0-387-70637-5 |
| Ocimum tenuiflorum | CASID:523-47-7 | beta-cadinene | ISBN:9788171360536 |
| Ocimum tenuiflorum | CID:10050 | (1S)-1,7,7-Trimethylbicyclo[2.2.1]heptan-2-one | ISBN:9788171360536 |
| Ocimum tenuiflorum | CID:10364 | CARVACROL | ISBN:978-0-387-70637-5, ISBN:9788171360536 |
| Ocimum tenuiflorum | CID:10655819 | gamma-Selinene | ISBN:9788171360536 |
| Ocimum tenuiflorum | CID:12223113 | 1S-.alpha.-Pinene | DOI:10.15482/USDA.ADC/1239279 |
| Ocimum tenuiflorum | CID:12303662 | Phytosterols | ISBN:9788171360536 |

CSV



B

**luteolin**



**Download:** SDF or MOL
**Pubchem Identifier:** 5280445
**ChEBI Identifier:** 15864
**CAS Identifier:** 491-70-3
**CHEMSPIDER identifier:** 4444102
**Synonyms:** LUTEOLIN

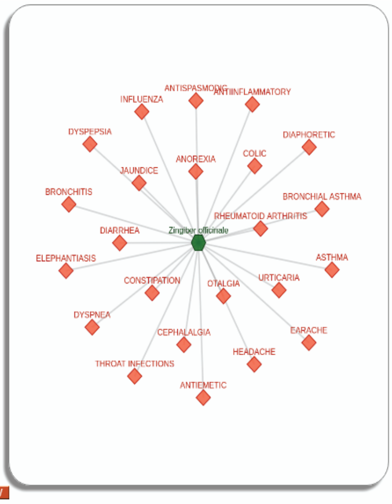
**Physicochemical Properties**

| | |
|---|---|
| Molecular weight | 286.24 g/mol |
| LogP | 2.53 |
| LogD | 1.56 |
| LogSw | -3.52 |
| Topological polar surface area | 110.80 Å$^2$ |
| Number of hydrogen bond donors | 4 |
| Number of hydrogen bond acceptors | 6 |

C

| INDIAN MEDICINAL PLANT | THERAPEUTIC USE | THERAPEUTIC IDENTIFIER | REFERENCES |
|---|---|---|---|
| Zingiber officinale | ANOREXIA | MESH:D000856<br>UMLS:C1847492<br>DOID:8689 | ISBN:978-0-387-70637-5, ISBN:9788171360536 |
| Zingiber officinale | ANTIEMETIC | | ISBN:978-0-387-70637-5 |
| Zingiber officinale | ANTIINFLAMMATORY | | ISBN:978-0-387-70637-5 |
| Zingiber officinale | ANTISPASMODIC | | ISBN:978-0-387-70637-5 |
| Zingiber officinale | ASTHMA | MESH:D001249, MESH:D001250, MESH:D016535<br>UMLS:C0004096, UMLS:C0004099, UMLS:C0085129, UMLS:C0155883<br>DOID:2841 | ISBN:9788171360536 |
| Zingiber officinale | BRONCHIAL ASTHMA | MESH:D001249, MESH:D001250, MESH:D016535<br>UMLS:C0004096<br>DOID:2841 | ISBN:9788171360536 |

CSV




D

Physicochemical properties | Druggability filters

| Molecular weight (g/mol) | LogP | Topological polar surface area (Å²) | Number of hydrogen bond acceptor |
|---|---|---|---|
| Is equal to | Is equal to | Is equal to | Is equal to |

| Number of hydrogen bond donors | Number of heavy atoms | Number of heteroatoms | Number of rigid bonds |
|---|---|---|---|
| Is equal to | Is equal to | Is equal to | Is equal to |

Number of rotatable bonds
Is equal to
Apply

Select any filter and click on Apply to see results

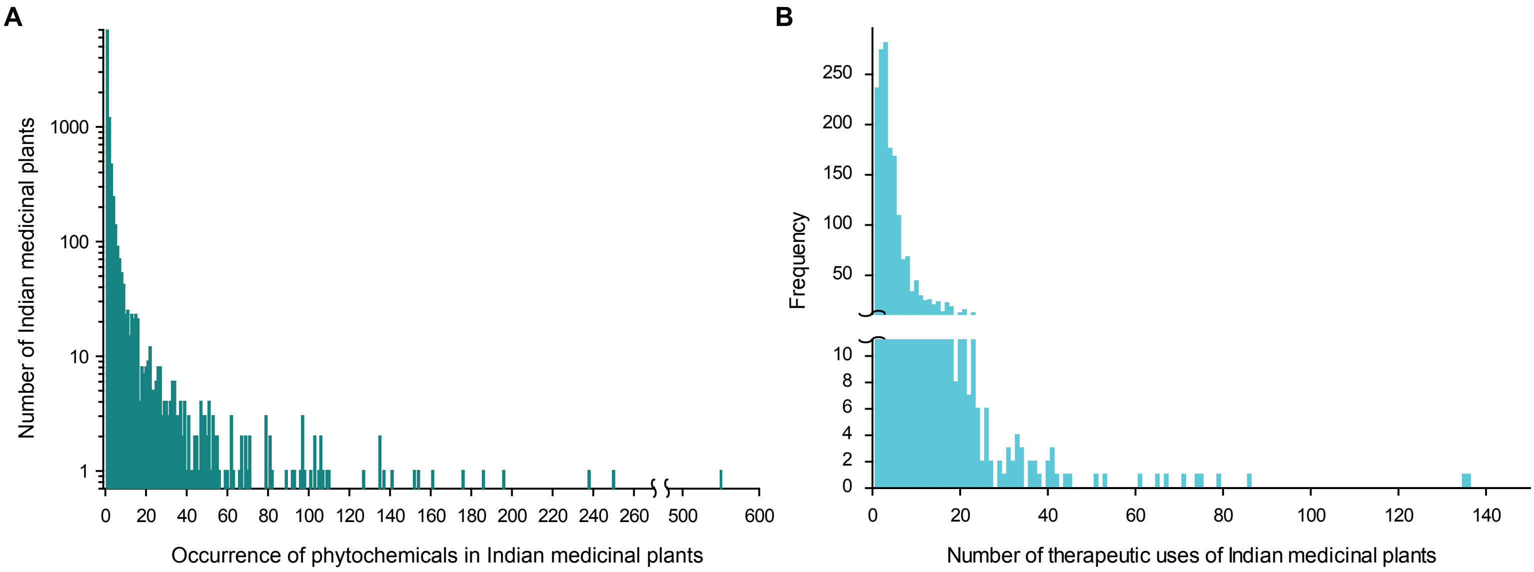
Physicochemical properties | Druggability filters

| Lipinski RO5 violation | Oral PhysChem score | GSK 4/400 | Pfizer 3/75 | Veber rule | Egan rule | QEDw |
|---|---|---|---|---|---|---|
| - Any - | - Any - | - Any - | - Any - | - Any - | - Any - | Is equal to |

Apply

Select any filter and click on Apply to see results

**Figure 2**



A
Number of Indian medicinal plants
1000
100
10
1
0
20
40
60
80
100
120
140
160
180
200
220
240
260
500
600
Occurrence of phytochemicals in Indian medicinal plants
B
Frequency
250
200
150
100
50
10
8
6
4
2
0
0
20
40
60
80
100
120
140
Number of therapeutic uses of Indian medicinal plants


**Figure 3**

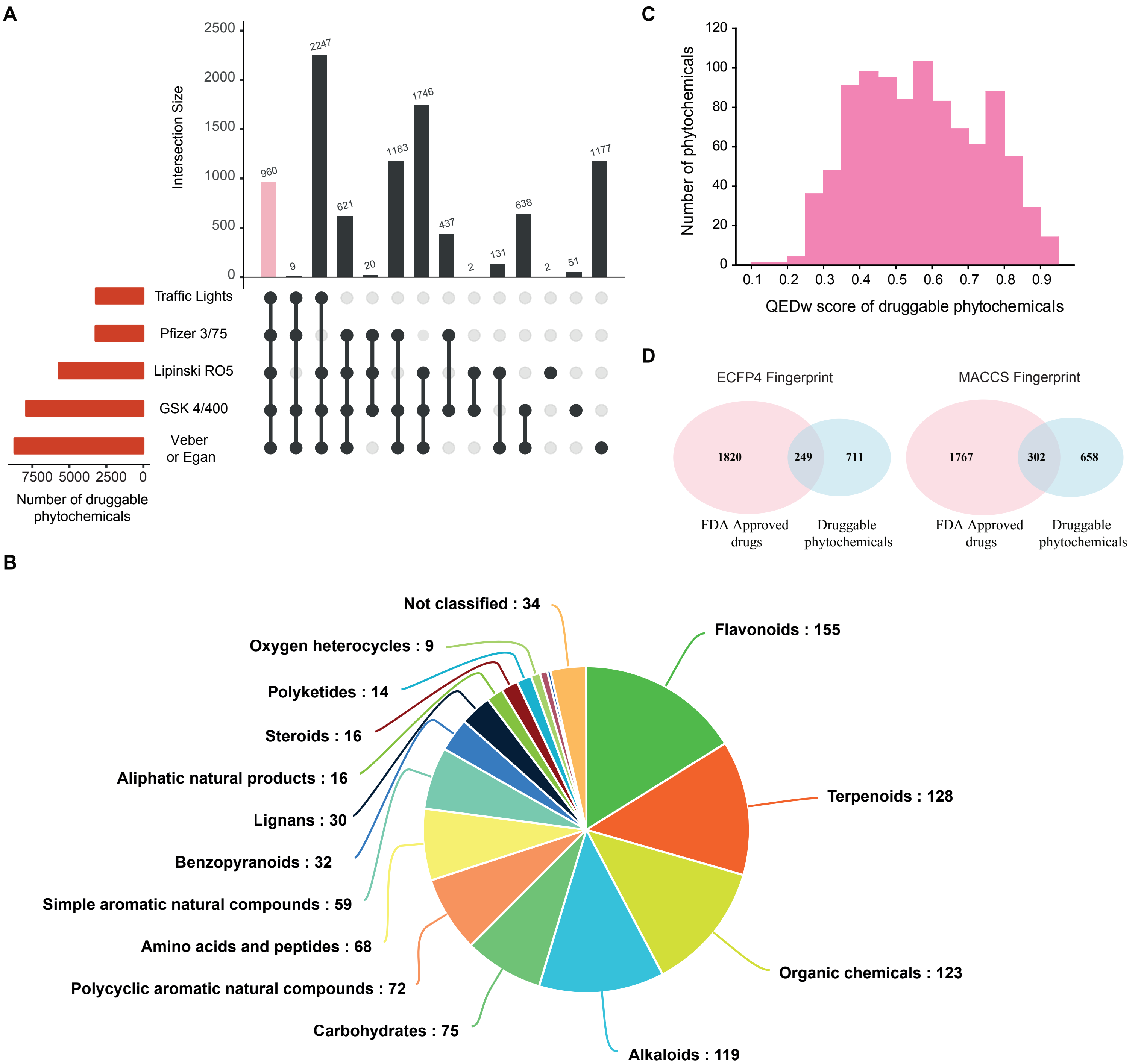
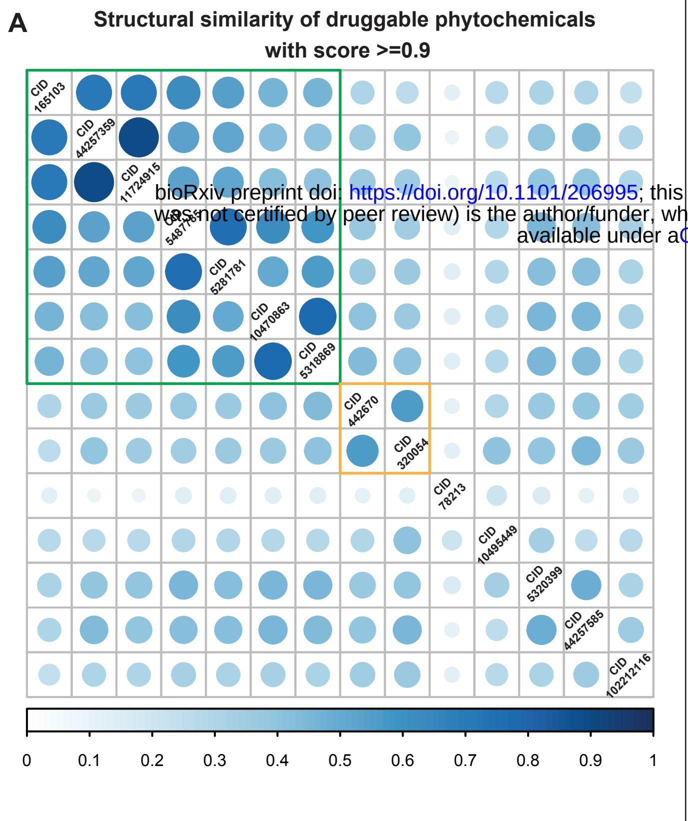
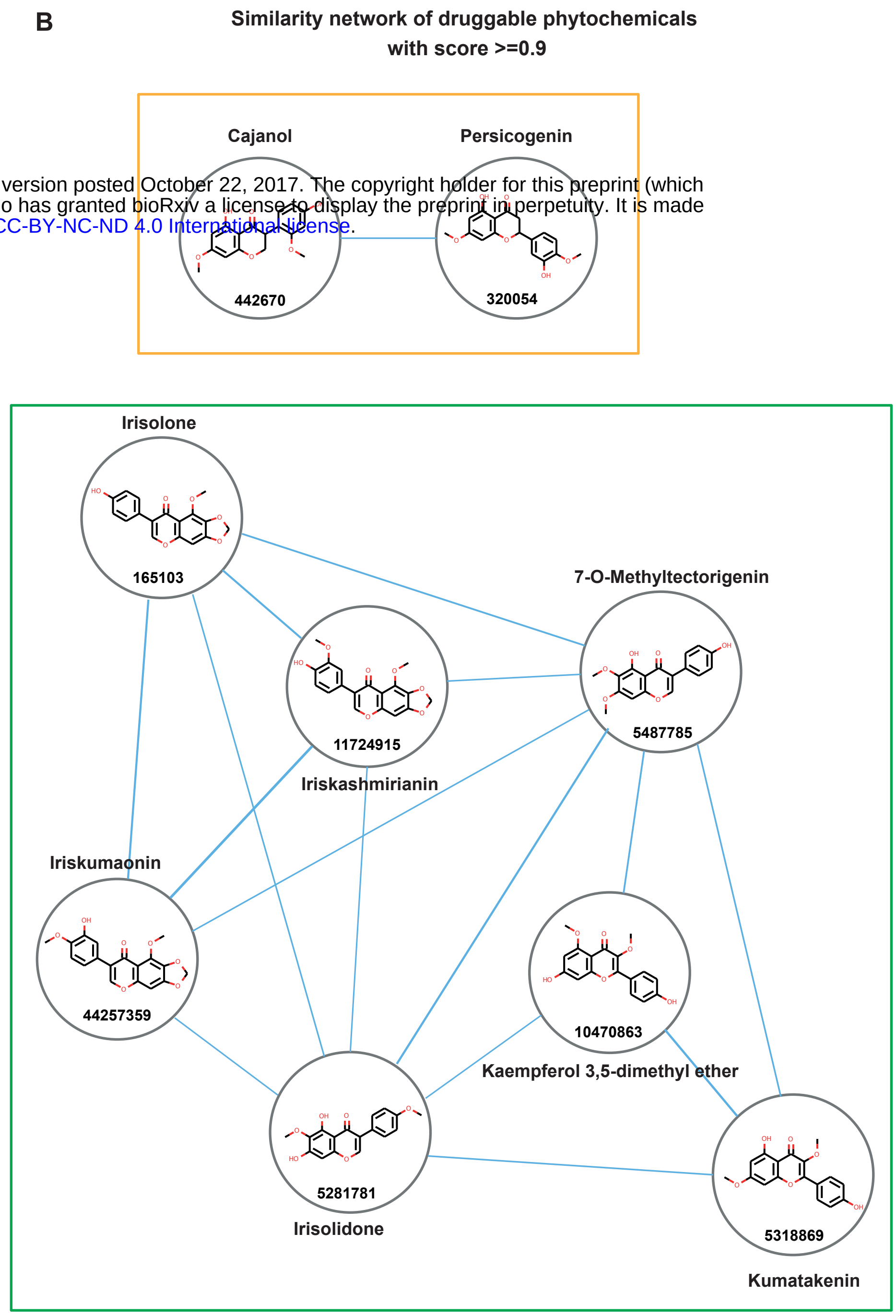
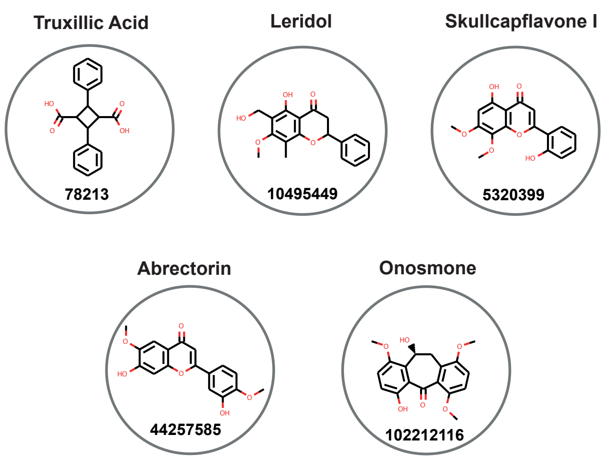

A
Intersection Size
2500
2000
1500
1000
500
0
960
9
2247
621
20
1183
1746
437
2
131
638
2
51
1177
Traffic Lights
Pfizer 3/75
Lipinski RO5
GSK 4/400
Veber or Egan
7500 5000 2500 0
Number of druggable phytochemicals
C
Number of phytochemicals
QEDw score of druggable phytochemicals
D
ECFP4 Fingerprint
1820
249
711
FDA Approved drugs
Druggable phytochemicals
MACCS Fingerprint
1767
302
658
FDA Approved drugs
Druggable phytochemicals
B
Not classified : 34
Flavonoids : 155
Oxygen heterocycles : 9
Polyketides : 14
Steroids : 16
Aliphatic natural products : 16
Lignans : 30
Benzopyranoids : 32
Simple aromatic natural compounds : 59
Amino acids and peptides : 68
Polycyclic aromatic natural compounds : 72
Carbohydrates : 75
Alkaloids : 119
Organic chemicals : 123
Terpenoids : 128

Figure 4


A
Structural similarity of druggable phytochemicals with score >=0.9
CID 165103
CID 44257359
CID 11724915
CID 5487785
CID 5281781
CID 10470863
CID 5318869
CID 442670
CID 320054
CID 78213
CID 10495449
CID 5320399
CID 44257585
CID 102212116
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1


B
Similarity network of druggable phytochemicals with score >=0.9
Cajanol
442670
Persicogenin
320054
Irisolone
165103
7-O-Methyltectorigenin
5487785
11724915
Iriskashmirianin
Iriskumaonin
44257359
10470863
Kaempferol 3,5-dimethyl ether
5281781
Irisolidone
5318869
Kumatakenin


Truxillic Acid
78213
Leridol
10495449
Skullcapflavone I
5320399
Abrectorin
44257585
Onosmone
102212116

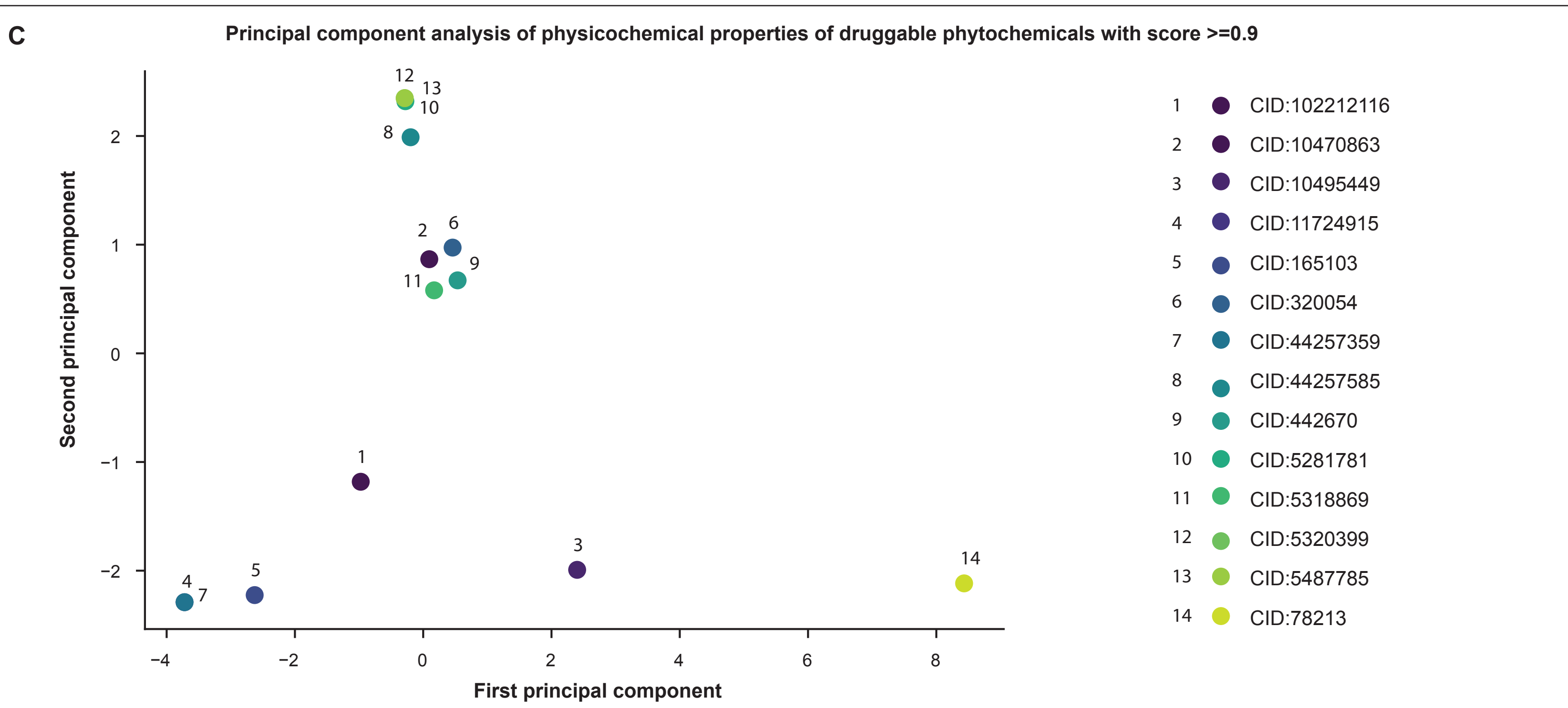

C
Principal component analysis of physicochemical properties of druggable phytochemicals with score >=0.9
Second principal component
First principal component
1 CID:102212116
2 CID:10470863
3 CID:10495449
4 CID:11724915
5 CID:165103
6 CID:320054
7 CID:44257359
8 CID:44257585
9 CID:442670
10 CID:5281781
11 CID:5318869
12 CID:5320399
13 CID:5487785
14 CID:78213

Figure 5

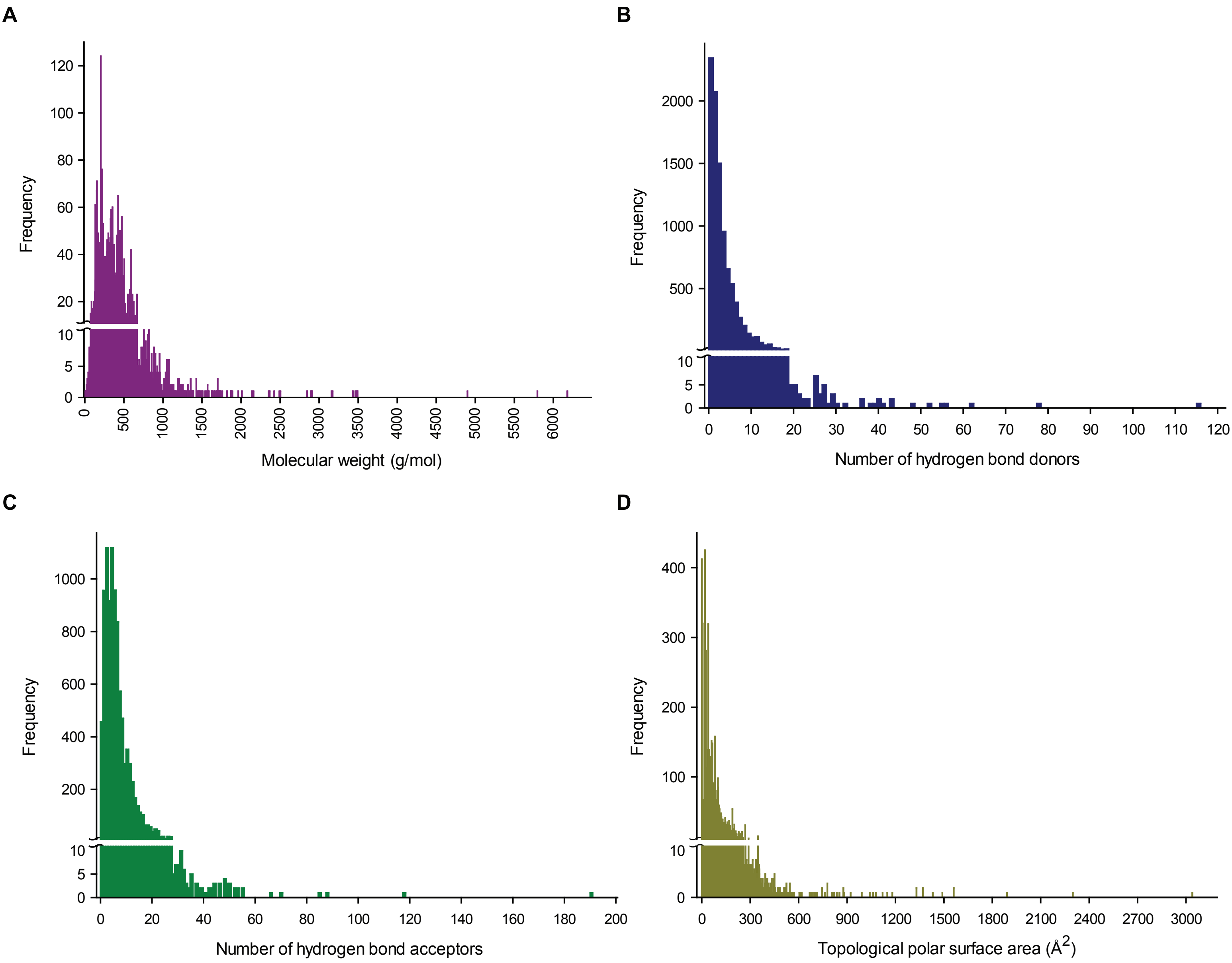

A
Frequency
Molecular weight (g/mol)
B
Frequency
Number of hydrogen bond donors
C
Frequency
Number of hydrogen bond acceptors
D
Frequency
Topological polar surface area (Å2)

**Figure 6**