IMPPAT 2.0: An Enhanced and Expanded Phytochemical Atlas of Indian Medicinal Plants

R.P. Vivek-Ananth, Karthikeyan Mohanraj, Ajaya Kumar Sahoo, and Areejit Samal*

ACCESS | 📊 Metrics & More | 📑 Article Recommendations | 🔲 Supporting Information

ABSTRACT: Compilation, curation, digitization, and exploration of the phytochemical space of Indian medicinal plants can expedite ongoing efforts toward natural product and traditional knowledge based drug discovery. To this end, we present IMPPAT 2.0, an enhanced and expanded database compiling manually curated information on 4010 Indian medicinal plants, 17,967 phytochemicals, and 1095 therapeutic uses. Notably, IMPPAT 2.0 compiles associations at the level of plant parts and provides a FAIR-compliant nonredundant *in silico* stereo-aware library of 17,967 phytochemicals from Indian medicinal plants. The phytochemical library has been annotated with several useful properties to enable easier exploration of the chemical space. We have also filtered a subset of 1335 drug-like phytochemicals of which majority have no similarity to existing approved drugs. Using cheminformatics, we have characterized the molecular complexity and molecular scaffold based structural diversity of the phytochemical space of Indian medicinal plants and performed a comparative analysis with other chemical libraries. Altogether, IMPPAT 2.0 is a manually curated extensive phytochemical atlas of Indian medicinal plants that is accessible at https://cb.imsc.res.in/imppat/.
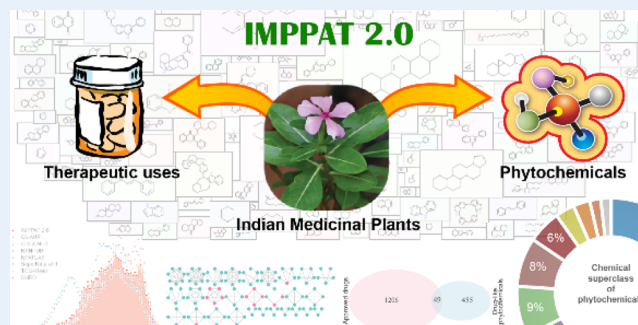
INTRODUCTION

Medicinal plants have been used for centuries to treat human ailments in different systems of traditional medicine across the world. Phytochemicals are the chemical factors behind the therapeutic action of such plants and the medicinal formulations prepared from them.[1,2] Consequently, significant research has been directed toward the identification of phytochemicals of medicinal plants[3−6] to discover novel and biologically relevant molecules. Furthermore, phytochemicals along with other natural products represent a biologically relevant chemical space produced by diverse organisms that have evolved to attain a high level of fitness under varied selective pressures.[7] These aspects have rendered the natural product space as a key player in the identification and development of drugs against several diseases. This fact is cemented by the recent analysis by Newman and Cragg[8] wherein the authors report that 34% of the approved small molecule drugs in the last four decades are natural products, natural product derived, or botanical drugs.[8] Yet, the pursuit for natural product based drug discovery has declined since 1990s because of several challenges including sourcing of biological materials, extraction and isolation of bioactive compounds, and the high structural complexity of the natural products.[9,10] For instance, Atanasov *et al.*[9] in a recent review discussed the current technological advances that can help overcome many of the above-mentioned challenges in natural product based drug discovery. Further, much of the natural product space remains largely unexplored, providing a significant scope for the identification of novel molecular scaffolds and fragments for the development of new drugs.[8]

Although majority of the characterized natural products are either therapeutic or nontoxic, several natural products have been identified to be toxic to humans.[11] Thus, a careful evaluation of the source and toxicity of the natural product is essential. Indian medicinal plants have been used for ages in traditional Indian systems of medicine like Ayurveda and Siddha to treat a variety of human diseases.[12] These medicinal plants that are a rich source of novel phytochemicals are more likely enriched with therapeutic natural products. Much of the traditional knowledge on Indian medicinal plants still largely remains buried in books and monographs. The advent of several computational methods including artificial intelligence has reinvigorated natural product based drug discovery.[13−15] In this context, the nondigital nature of the information on traditional knowledge on Indian medicinal plants limits their complete and effective use in drug discovery research. Further, the molecular mechanisms behind the therapeutic action of medicinal plants used in traditional Indian medicine remain
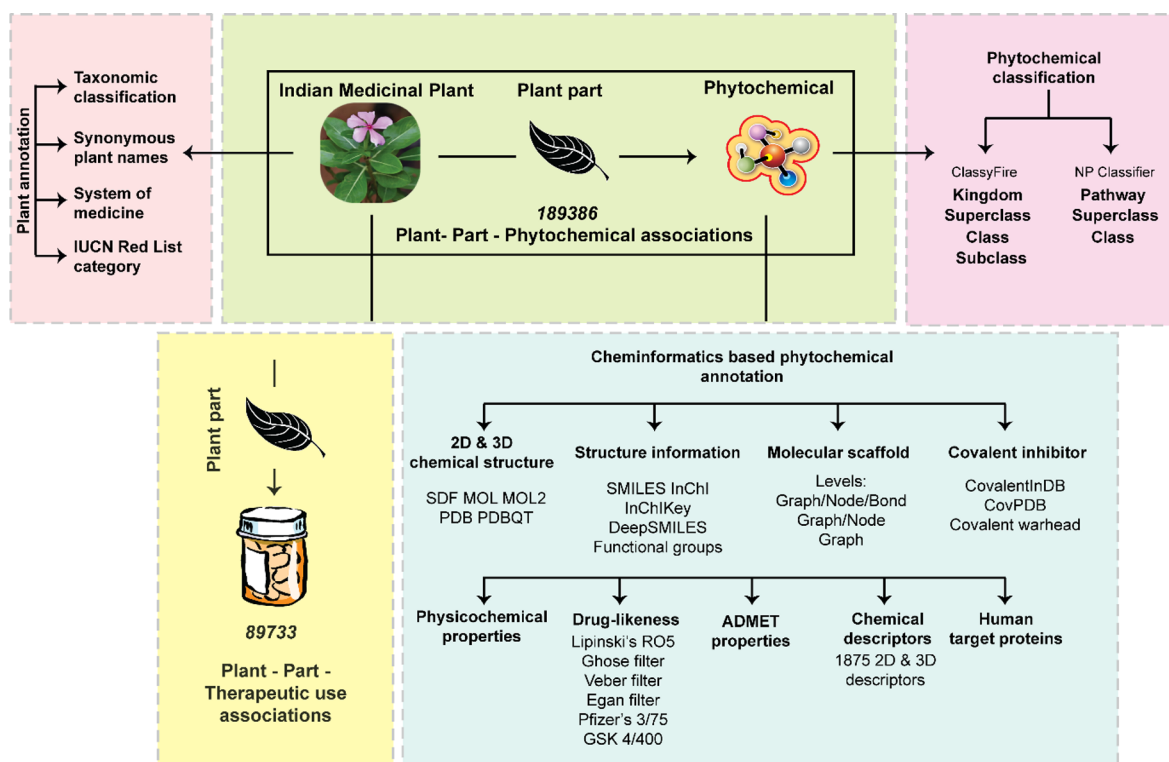
Figure 1. Schematic overview of the important features including enhancements and expansion realized in IMPPAT 2.0.

largely undiscovered. This poses a significant challenge toward turning a largely experience-based enterprise to evidence-based practice, leading to the modernization of traditional Indian medicine. In a nutshell, creation of a comprehensive database on Indian medicinal plants, their phytochemicals, and their therapeutic uses will be of immense use in natural product and traditional knowledge based drug discovery.

Toward this goal, we had earlier built the manually curated database IMPPAT (version 1.0)[16] containing 1742 Indian Medicinal Plants, their 9596 Phytochemicals, And their Therapeutic uses. Importantly, IMPPAT 1.0 compiled two-dimensional (2D) and three-dimensional (3D) chemical structures of the 9596 phytochemicals in the database, along with their physicochemical; drug-likeness; and absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties. In short, IMPPAT 1.0 provided a large phytochemical atlas specific to Indian medicinal plants.[16,17] Subsequent to publication, the IMPPAT phytochemical library has enabled several computer-aided drug discovery studies, including research on the identification of anti-SARS-CoV-2 drugs.[18−23]

Here we present IMPPAT 2.0, an enhanced and expanded phytochemical atlas of Indian medicinal plants (Figure 1). The latest update, IMPPAT 2.0, has built upon the published data of the earlier version[16] and now compiles information on 4010 Indian medicinal plants, 17,967 phytochemicals, and 1095 therapeutic uses (Table 1). We have highlighted the key features of IMPPAT 2.0 in Figure 1. In IMPPAT 2.0, the coverage of the Indian medicinal plants is more than doubled, and the phytochemical and therapeutic use associations of the Indian medicinal plants have increased more than 5-fold in comparison with IMPPAT 1.0. Also, IMPPAT 2.0 now provides the phytochemical composition and therapeutic uses of Indian medicinal plants at the level of plant parts

Table 1. Comparison of the Updated Version (IMPPAT 2.0) with the Previous Version (IMPPAT 1.0)

feature	IMPPAT 2.0	IMPPAT 1.0
number of Indian medicinal plants	4010	1742
number of phytochemicals	17,967	9596
number of plant−part−phytochemical associations	189,386	*not available*
number of plant−phytochemical associations	124,995	27,074
number of therapeutic uses	1095	1124
number of plant−part−therapeutic use associations	89,733	*not available*
number of plant−therapeutic use associations	60,732	11,514

such as the stem, root, or leaves. Further, through extensive manual curation and standardization, IMPPAT 2.0 provides a FAIR[24]-compliant nonredundant *in silico* stereo-aware library of 17,967 phytochemicals with 2D and 3D chemical structures. Subsequently, we have characterized the molecular complexity and the molecular scaffold based structural diversity of the phytochemical space of IMPPAT 2.0 and, thereafter, compared these with those of other chemical libraries. We have also filtered a subset of 1335 drug-like phytochemicals using multiple drug-likeness rules. Finally, we have compared the phytochemicals in IMPPAT 2.0 with phytochemicals from Chinese medicinal plants. From our cheminformatics analysis, we highlight the uniqueness, utility, and complementary nature of the phytochemical space of Indian medicinal plants captured in IMPPAT 2.0. IMPPAT 2.0 is among the published natural product databases characterizing the natural products specific to geographical regions such as AfroDb,[25] TCM-Mesh,[26] and several Latin American natural product databases.[27] IMPPAT 2.0 is accessible without any login or registration requirement
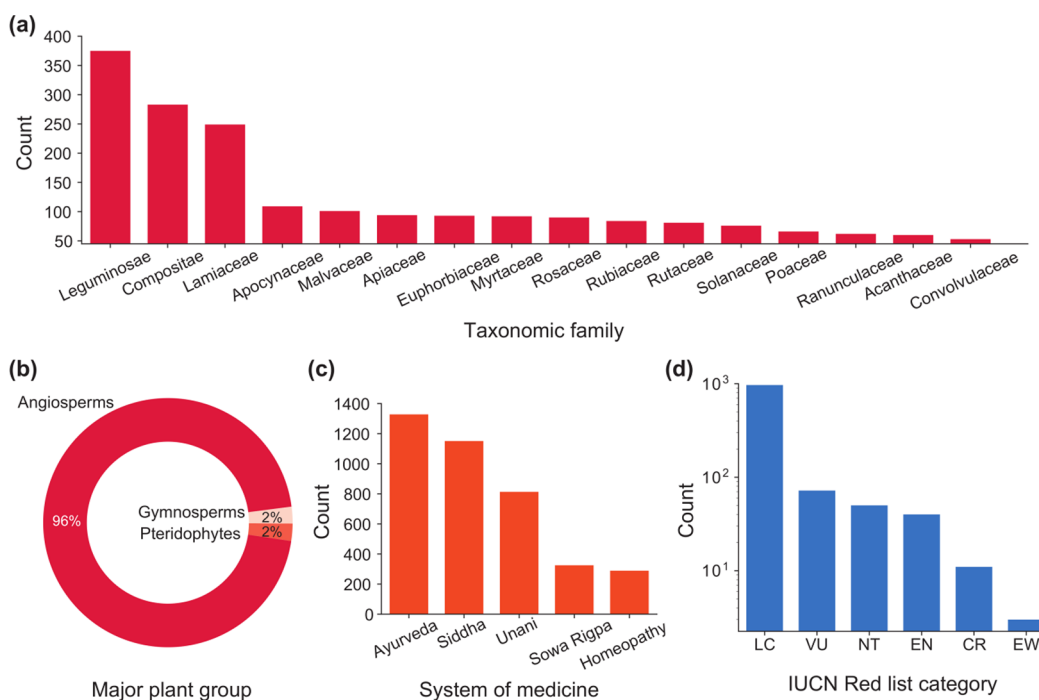
Figure 2. Coverage of Indian medicinal plants in IMPPAT 2.0. (a) Top taxonomic families of Indian medicinal plants in IMPPAT 2.0. Note that only families with more than 50 Indian medicinal plants in the database are shown. (b) Classification of the Indian medicinal plants in IMPPAT 2.0 into major plant groups: Angiosperms (flowering plants), Gymnosperms (conifers and cycads), and Pteridophytes (ferns and fern allies). (c) Use of Indian medicinal plants in traditional Indian systems of medicine such as Ayurveda, Siddha, Unani, Sowa-Rigpa, and Homeopathy. Note that a given Indian medicinal plant can be used in multiple systems of medicine. (d) Present category of the Indian medicinal plants in IMPPAT 2.0 according to conservation status based on the IUCN Red list of threatened species. LC, least concern; VU, vulnerable; NT, near threatened; EN, endangered; CR, critically endangered; and EW, extinct in the wild.

via a user-friendly web interface at https://cb.imsc.res.in/imppat/.

RESULTS AND DISCUSSION

Enhancement and Expansion of IMPPAT. The previous version, IMPPAT 1.0,[16] released in January 2018 was the largest online resource on phytochemicals of Indian medicinal plants. Here, we present the updated version, IMPPAT 2.0, which is a significant enhancement and expansion over the previous version (Table 1). This update was realized through extensive manual curation and addition of several new features to IMPPAT (Figure 1, Table 1). Figure 1 summarizes the important features including enhancements accomplished in IMPPAT 2.0.

Increase in Coverage of Indian Medicinal Plants. IMPPAT 2.0 compiles curated information on phytochemicals and therapeutic uses of 4010 Indian medicinal plants. The updated database achieves a more than 2-fold increase in the coverage of medicinal plants with respect to the previous version (Table 1). During data collection from various sources, we encountered extensive use of synonymous plant names in the published literature reporting information on phytochemicals and therapeutic uses of medicinal plants. This use of synonymous plant names can create difficulties while choosing the correct plant for phytochemical extraction or preparation of pharmaceutical formulations as prescribed in traditional medicine pharmacopeia. For this reason, IMPPAT 2.0 provides the compiled information for a nonredundant list of 4010 Indian medicinal plants. This nonredundant list was created via an extensive manual curation effort as follows. First, we compiled a list of more than 7000 synonymous names

corresponding to Indian medicinal plants for which the phytochemical information was collected from the published literature in IMPPAT 1.0 or during this update. Second, The Plant List database[28] was used to identify the accepted scientific names for the compiled plant names. Third, the synonymous names were merged using the accepted scientific names.

Further, the Indian medicinal plants covered in IMPPAT 2.0 have been annotated with information on their taxonomic classification, their use in traditional Indian systems of medicine, their synonymous names, and their present category in the IUCN Red list of threatened species[29] (Methods). The 4010 Indian medicinal plants in IMPPAT 2.0 belong to 244 taxonomic families, and Figure 2a shows the families with more than 50 Indian medicinal plants in our database. In particular, Leguminosae is the largest family with more than 350 plants in IMPPAT 2.0. This is expected as Leguminosae, commonly known as legume, pea, or bean family, is a large and medicinally important family of flowering plants.[30] The next two large families in IMPPAT 2.0 are Compositae and Lamiaceae, both of which are again families of flowering plants. Flowering plants or Angiosperms constitute 96% of the plants in IMPPAT 2.0. The remaining plants are Gymnosperms (2%), which include conifers and cycads, and Pteridophytes (2%), which include ferns and fern allies (Figure 2b). The medicinal plants captured in IMPPAT 2.0 are used in one or more traditional Indian systems of medicine such as Ayurveda, Siddha, Unani, Sowa-Rigpa, and Homeopathy. In particular, 1328 plants in IMPPAT 2.0 are used in Ayurveda followed by 1151 plants used in Siddha (Figure 2c). Precariously, we find that many of the Indian medicinal plants require extensive
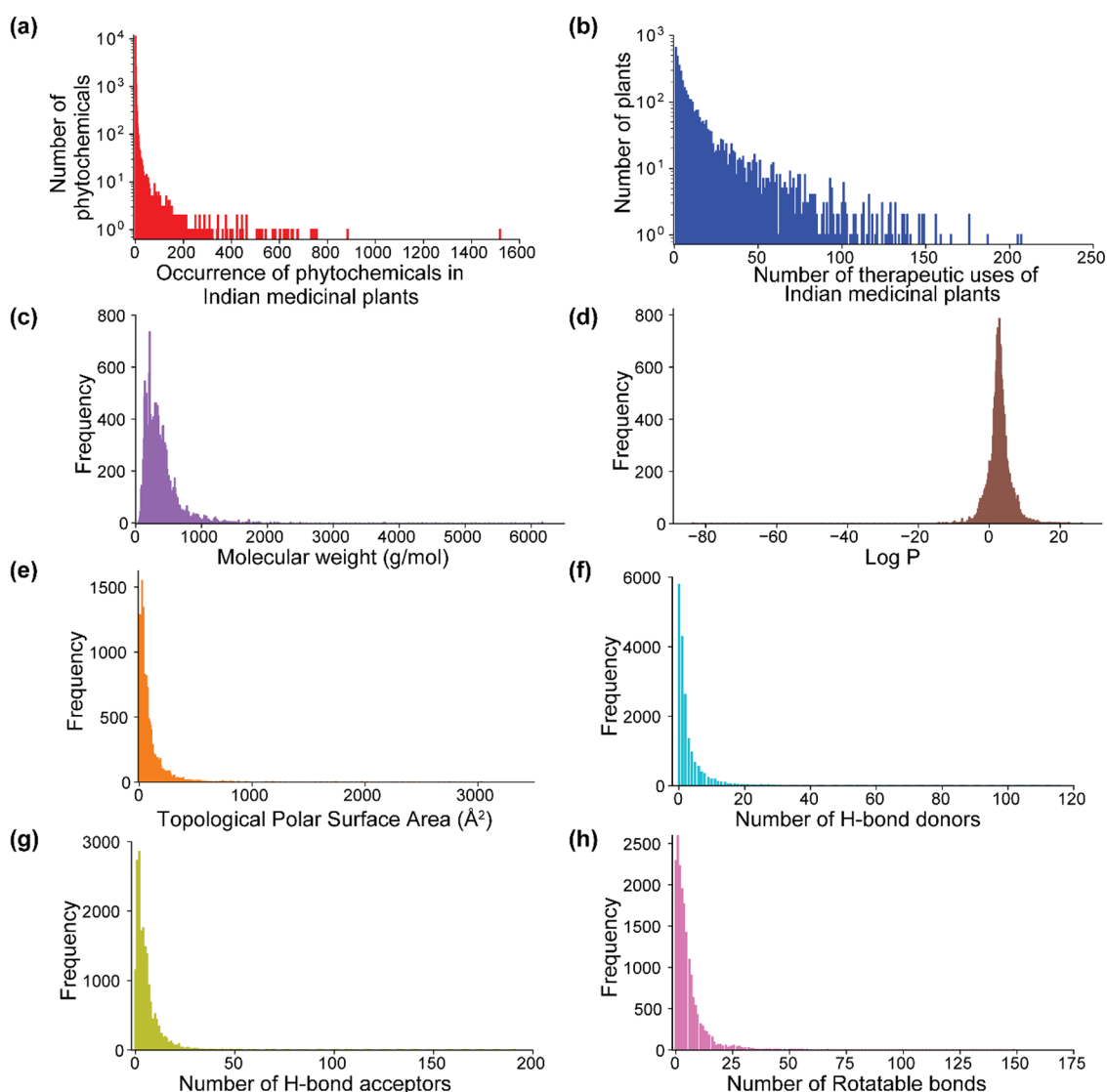
Figure 3. Basic statistics and distribution of the physicochemical properties for phytochemicals in IMPPAT 2.0. (a) Histogram of the number of Indian medicinal plants that produce a given phytochemical in IMPPAT 2.0. (b) Histogram of the number of therapeutic uses per Indian medicinal plant in IMPPAT 2.0. Distribution of six important physicochemical properties for 17,967 phytochemicals, namely, (c) molecular weight (g/mol), (d) log P, (e) topological polar surface area ($Å^2$), (f) number of hydrogen bond (H-bond) donors, (g) number of hydrogen bond (H-bond) acceptors, and (h) number of rotatable bonds.

conservation efforts as 72, 50, 40, 11, and 3 plants are categorized in the IUCN Red list of threatened species[29] as vulnerable (VU), near threatened (NT), endangered (EN), critically endangered (CR), and extinct in the wild (EW), respectively (Figure 2d).

Information at the Level of Plant Parts. Another significant update over version 1.0 is that IMPPAT 2.0 now provides information on plant−phytochemical and plant−therapeutic use associations at the level of plant parts (Table 1). For instance, the updated database compiles published information on the phytochemical composition for any Indian medicinal plant at the level of plant parts such as the stem, root, or leaves. Because the phytochemical composition can significantly vary across different plant parts, this enhancement in IMPPAT 2.0 will enable phytochemists and pharmacognosists to choose the appropriate protocol for extraction of the phytochemical of their interest for drug discovery studies. Moreover, traditional Indian systems of medicine, such as Ayurveda and Siddha, use specific plant parts to treat various

diseases. Thus, the compiled information in IMPPAT 2.0 on therapeutic use at the level of plant parts will help in better understanding the therapeutic potential of the Indian medicinal plants.

Increase in Coverage of Phytochemicals. A major enhancement in IMPPAT 2.0 is the creation of a non-redundant stereo-aware natural product library of 17,967 phytochemicals specific to Indian medicinal plants. This represents a nearly 2-fold expansion in the size of the phytochemical library in comparison to the previous version (Table 1).

Building upon the published methodology and extensive data compiled in IMPPAT 1.0,[16] we expanded the phytochemical associations in IMPPAT 2.0 as follows. First, the bulk of the plant−part−phytochemical associations for Indian medicinal plants were manually collected, curated, and digitized from 70 specialized books (Table S1). Only 9 out of these 70 books were covered in IMPPAT 1.0. Importantly, the remaining 61 books covered in IMPPAT 2.0 include the (a) 5
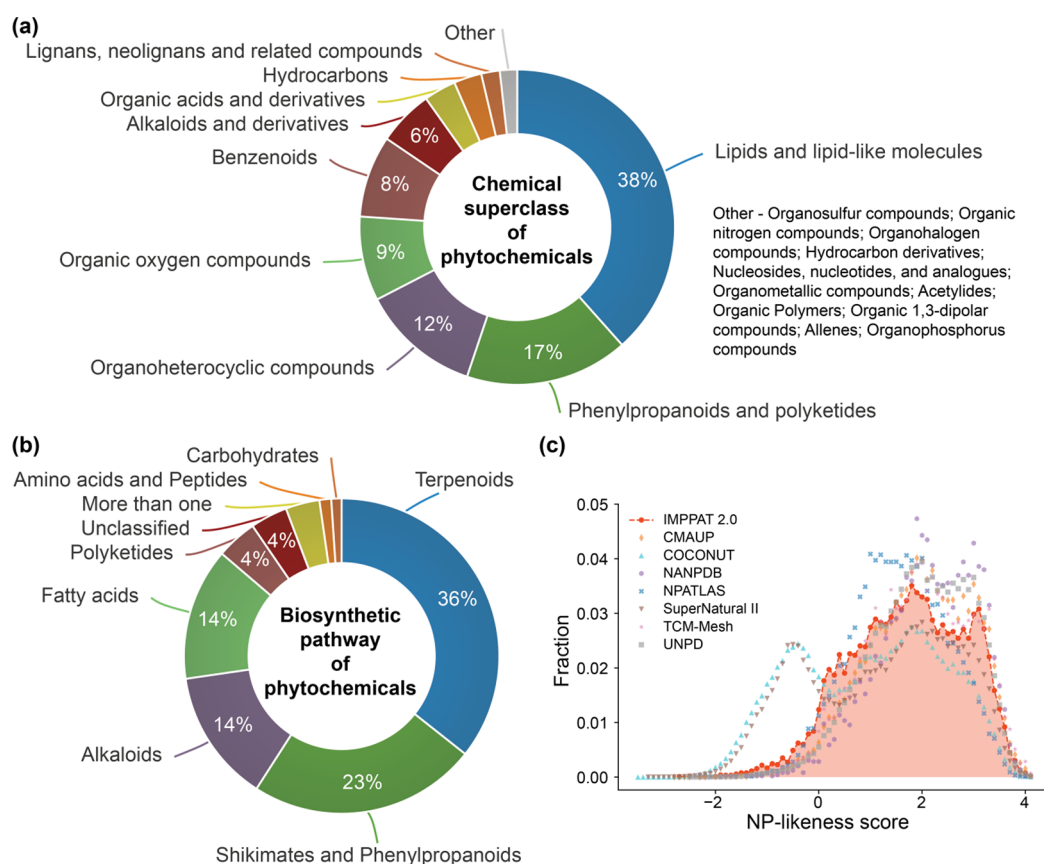
Figure 4. Chemical classification, biosynthetic pathways, and natural product likeness of phytochemicals in IMPPAT 2.0. (a) Chemical superclasses of phytochemicals predicted by ClassyFire.[40] (b) Biosynthetic pathways for phytochemicals predicted by NP classifier.[38] (c) Distribution of the NP-likeness scores for phytochemicals in IMPPAT 2.0 and other natural product libraries.

volumes of The Wealth of India published by the Council of Scientific and Industrial Research, Government of India; (b) 14 volumes of Ayurvedic, Siddha, and Unani pharmacopeias of India published by the Ministry of AYUSH, Government of India; and (c) 18 volumes of the Reviews of Indian medicinal plants published by the Indian Council of Medical Research (ICMR), Government of India. These valuable yet non-digitized book sources on Indian medicinal plants are known for their comprehensiveness and accuracy.[31] Second, aside from the books, all the plant–phytochemical associations compiled from various sources in the previous version (IMPPAT 1.0)[16] were manually revisited to additionally gather and curate phytochemical information at the level of plant parts. This last step also involved manual curation of more than 7000 research articles covered in IMPPAT 1.0 to gather additional information at the level of plant parts. Third, we incorporated data from a published database[32] providing phytochemical information for the Indian medicinal plant *Rauvolfia serpentina*.

A major challenge during the compilation, curation, and digitization of the phytochemical composition of Indian medicinal plants is the large-scale use of nonstandard and synonymous names for phytochemicals in books and research articles. Therefore, to create a nonredundant list of phytochemicals, we have standardized the phytochemical names fetched from diverse sources as follows. First, we mapped the chemical names to identifiers in standard databases such as PubChem[33] and retrieved the associated two-dimensional (2D) and three-dimensional (3D) structures.

Second, we compared the phytochemicals based on their structural similarity. Third, we manually checked the stereo-chemistry of the phytochemicals using the InChI. These steps led to the creation of a nonredundant stereo-aware chemical library of 17,967 phytochemicals that are produced by 4010 Indian medicinal plants with therapeutic uses. Thus, the phytochemical atlas will aid ongoing efforts toward the identification of novel bioactive and therapeutic molecules.

Overall, there are 189,386 nonredundant plant–part–phytochemical associations in IMPPAT 2.0 spanning 4010 Indian medicinal plants and 17,967 phytochemicals. At the level of plant–phytochemical associations (after ignoring the plant parts), there is a 5-fold increase in IMPPAT 2.0 (Table 1). Figure 3a shows the occurrence of phytochemicals across 4010 Indian medicinal plants in IMPPAT 2.0. It can be seen that a majority of the phytochemicals (15335) have been reported to be produced by <5 Indian medicinal plants, whereas a minority of the phytochemicals (114) are produced by >200 Indian medicinal plants. In IMPPAT 2.0, *Psidium guajava* (468), *Citrus sinensis* (457), *Catharanthus roseus* (427), *Coriandrum sativum* (403), *Artemisia annua* (393), *Rosmarinus officinalis* (391), *Daucus carota* (391), *Origanum vulgare* (366), *Citrus reticulata* (364), and *Salvia officinalis* (363) are the top 10 plants in terms of the compiled information on the number of phytochemicals produced by them.

Enhanced Annotation to Enable Exploration of the Phytochemical Space. We have significantly enhanced the additional information on phytochemicals in IMPPAT 2.0, and

we now describe some of these new features in the updated database.

To make the phytochemical library of IMPPAT 2.0 compliant with Findable, Accessible, Interoperable, and Reusable (FAIR) principles, we assign unique IMPPAT identifiers to phytochemicals in the database, and thereafter, the identifiers are annotated with chemical names, structural features, and external links to standard chemical databases. Moreover, we provide the 2D and 3D chemical structures of phytochemicals in five different file formats (Methods).

The molecular scaffold represents the core structure of a molecule and is a key concept with wide applications in medicinal chemistry. In IMPPAT 2.0, we used the definition by Lipkus et al.[34,35] to compute and provide the molecular scaffolds for phytochemicals at three levels (Methods). This scaffold information can be used by a chemist to group and retrieve phytochemicals with the same core structure to further build upon them. In IMPPAT 2.0, we also used the definition by Ertl[36] to provide the functional groups present in phytochemicals. This functional group information can also facilitate the exploration of the phytochemical space by chemists. Further enhancements of phytochemical annotation in IMPPAT 2.0 include new information such as Deep-SMILES,[37] which is an adaptation of SMILES for use in machine learning; natural product specific chemical classification from NP classifier;[38] and natural product likeness (NP-likeness)[39] score.

Molecular descriptors capture important structural features and are useful in machine learning based classification and regression analysis such as Quantitative Structure Activity Relationship (QSAR). In IMPPAT 2.0, we provide 1875 2D and 3D chemical descriptors for each phytochemical. Lastly, drug-likeness scores can enable selection of chemicals with favorable properties as drug lead molecules. In IMPPAT 2.0, we also evaluated the drug-likeness of phytochemicals based on multiple scores computed using in-house scripts (Methods).

Figure 3c–h shows the distribution of six important physicochemical properties for the 17,967 phytochemicals in IMPPAT 2.0. On the basis of the chemical classification obtained by ClassyFire,[40] the 17,967 phytochemicals have been hierarchically categorized into 20 superclasses, 250 classes, and 410 subclasses. Among the 20 superclasses, Lipids and lipid-like molecules, Phenylpropanoids and polyketides, and Organoheterocyclic compounds are the top three with 6904, 3007, and 2202 phytochemicals, respectively (Figure 4a). Further, using NP classifier,[38] the 17,967 phytochemicals have been classified into one of seven biosynthetic pathways for natural products. Terpenoids, Shikimates and Phenylpropanoids, and Alkaloids are the top three biosynthetic pathways with 6049, 4206, and 2446 phytochemicals, respectively (Figure 4b).

The NP-likeness[39] score is a measure to quantify the natural product likeness of a given chemical structure. This score ranges from −5 to 5; the higher the score is, more likely that the molecule is a natural product.[41] Previous studies have shown that the NP-likeness of natural product libraries is predominantly positive and, moreover, is different from that of synthetic libraries that is predominantly negative.[42,43] On expected lines, phytochemicals in IMPPAT 2.0 have a predominantly positive NP-likeness score (>93%). Further, the distribution of the NP-likeness scores for phytochemicals in IMPPAT 2.0 is found to be similar to other natural product libraries (Figure 4c).

Lastly, IMPPAT 2.0 compiles information on 27,365 predicted interactions between phytochemicals and human target proteins from the STITCH[44] database. These 27,365 interactions involve 1294 phytochemicals and 5042 human target proteins.

Increase in Coverage of Therapeutic Uses. Building upon the compiled information in IMPPAT 1.0,[16] we enhanced the therapeutic use information in IMPPAT 2.0 to the level of plant parts and expanded it to cover the 4010 Indian medicinal plants in the updated database. This information on therapeutic use of Indian medicinal plants was compiled from 146 books on traditional medicine (Table S2). Only 9 out of these 146 books were covered in IMPPAT 1.0. Further, there are 56 books common to the set of 70 books from which phytochemical information was compiled and the set of 146 books from which therapeutic use information was compiled (Tables S1 and S2).

Because the therapeutic use of medicinal plants is reported using synonymous terms across different books, we undertook a manual curation effort to standardize the therapeutic use terms in IMPPAT 2.0. Specifically, we mapped the therapeutic use terms compiled from different books to standardized terms from Medical Subject Headings (MeSH),[45] International Classification of Diseases 11th Revision (ICD-11),[46] Unified Medical Language System (UMLS),[47] and Disease Ontology.[48] In the end, this effort to map the ethnopharmacological information on Indian medicinal plants to the standard vocabulary used in modern medicine led to a nonredundant list of 1095 standardized therapeutic use terms in IMPPAT 2.0.

Overall, there are 89,733 nonredundant plant−part−therapeutic use associations in IMPPAT 2.0 spanning 4010 Indian medicinal plants and 1095 standardized therapeutic uses. At the level of plant−therapeutic use associations (after ignoring the plant parts), there is a 5-fold increase in IMPPAT 2.0 (Table 1). Figure 3b shows the histogram of the number of therapeutic uses per Indian medicinal plant in IMPPAT 2.0. Whereas 21% of the Indian medicinal plants (851) in IMPPAT 2.0 have >20 therapeutic uses, the majority of Indian medicinal plants (2488) have <10 therapeutic uses.

Web Design and Data Accessibility. The webserver for the previous version (IMPPAT 1.0) enabled users to easily access the compiled information on Indian medicinal plants. Also, the IMPPAT 1.0 webserver enabled cheminformatics analysis such as filtering phytochemicals based on their physicochemical properties, drug-likeness scores, and chemical similarity. For the latest release (IMPPAT 2.0), we have completely redesigned the website. While incorporating all the features of the previous version, the web interface of IMPPAT 2.0 has multiple new features to facilitate the ease of use and exploration of the phytochemical space of Indian medicinal plants. This section describes some of the salient features of the IMPPAT 2.0 website. Users can access the compiled information in IMPPAT 2.0 via its web interface by three means, namely, browse, basic search, and advanced search.

Browse. In the web interface, users can browse the compiled information via the (a) phytochemical association and (b) therapeutic use association sections.

The phytochemical association section within browse enables a user to choose an Indian medicinal plant, a phytochemical, or a chemical superclass of phytochemicals to retrieve compiled information in IMPPAT 2.0 on plant−part−phytochemical associations along with literature references. If a specific plant is chosen, the user is redirected to a new page containing plant-
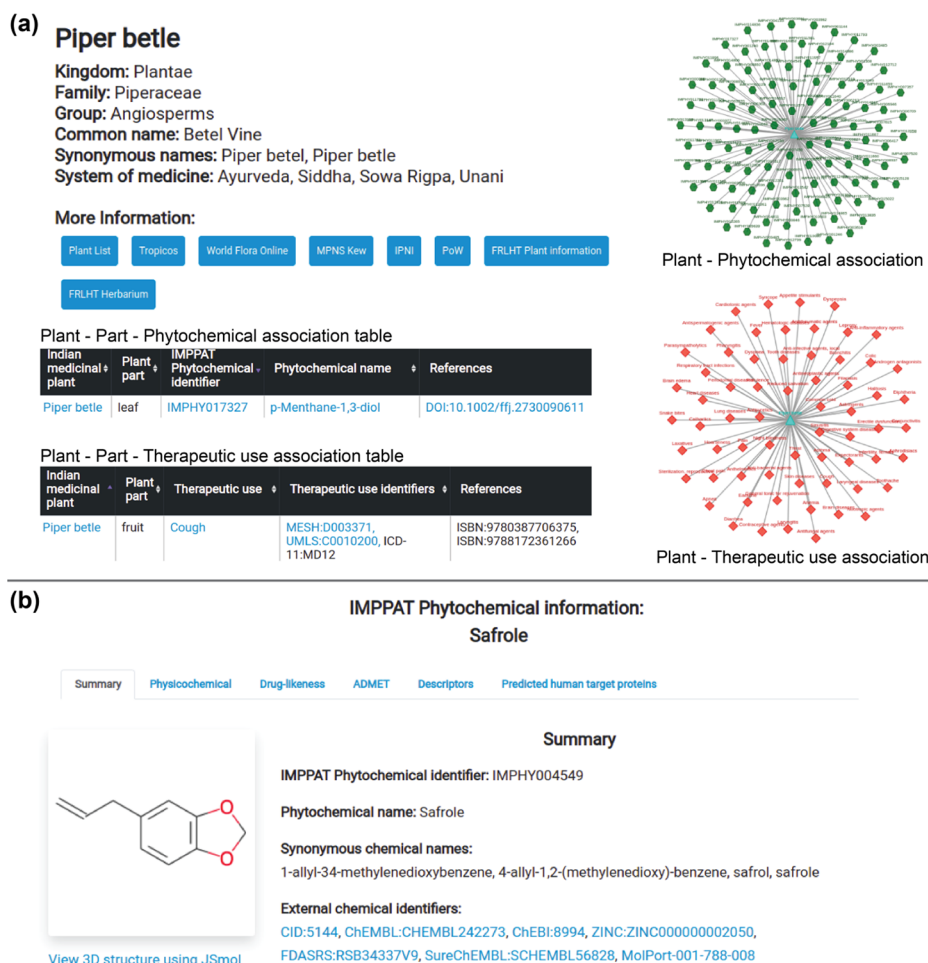
Figure 5. The web interface of the IMPPAT 2.0 database. (a) Snapshots of the results of queries for a phytochemical or a therapeutic use of an Indian medicinal plant. In this example, we show from IMPPAT 2.0 the snapshots of the plant information, plant−part−phytochemical association table, plant−part−therapeutic use association table, and network visualization of plant−phytochemical associations and plant−therapeutic use associations for *Piper betle*. (b) Screenshot of the dedicated page containing detailed information for the phytochemical Safrole.

specific information along with a table listing the phytochemical constituents for the plant at the level of plant parts (Figure 5a). The page also displays a network visualization of the plant−phytochemical associations, enabling the user to visually explore the phytochemical space of the chosen plant. If, instead of choosing a specific plant in the phytochemical association section, the user chooses a phytochemical or a chemical superclass of phytochemicals, the user is redirected to a new page containing a table listing the plant−part−phytochemical associations for the chosen phytochemical or for the phytochemicals belonging to the chosen chemical superclass.

Similar to the phytochemical association section within *browse*, the therapeutic use association section enables users to retrieve compiled information in IMPPAT 2.0 on the plant−part−therapeutic use associations with literature references by choosing either an Indian medicinal plant or a therapeutic use term (Figure 5a).

Basic Search. In the web interface, users can perform text-based searches in the *basic search* section to retrieve compiled information. The *basic search* section has two tabs: (a) phytochemical association and (b) therapeutic use.

In the phytochemical association tab, a user can perform text-based search using the complete or partial name of the plant, IMPPAT phytochemical identifier, or complete or partial name of the phytochemical to retrieve compiled information

on plant−part−phytochemical associations in IMPPAT 2.0. Upon submitting the text query, the user is presented with a table on the same page listing the relevant plant−part−phytochemical associations with literature references. In this table, the user can click any phytochemical name or identifier to view the page with detailed information on the phytochemical.

Similarly, in the therapeutic use tab, a user can perform text-based search using the complete or partial name of the plant or the therapeutic use term to retrieve compiled information on plant−part−therapeutic use associations in IMPPAT 2.0 with literature references as a table on the same page.

Advanced Search. In the web interface, the *advanced search* section enables a user to filter and retrieve a subset of phytochemicals compiled in IMPPAT 2.0 based on their physicochemical properties, drug-likeness, chemical similarity, and molecular scaffolds. The physicochemical filter tab provides a user with the option to retrieve phytochemicals of interest based on molecular weight, log P, topological polar surface area, hydrogen bond acceptors, hydrogen bond donors, number of heavy atoms, number of heteroatoms, number of rings, number of rotatable bonds, stereochemical complexity, and shape complexity. Similarly, the drug-like filter tab enables a user to filter phytochemicals based on multiple drug-likeness scoring schemes.
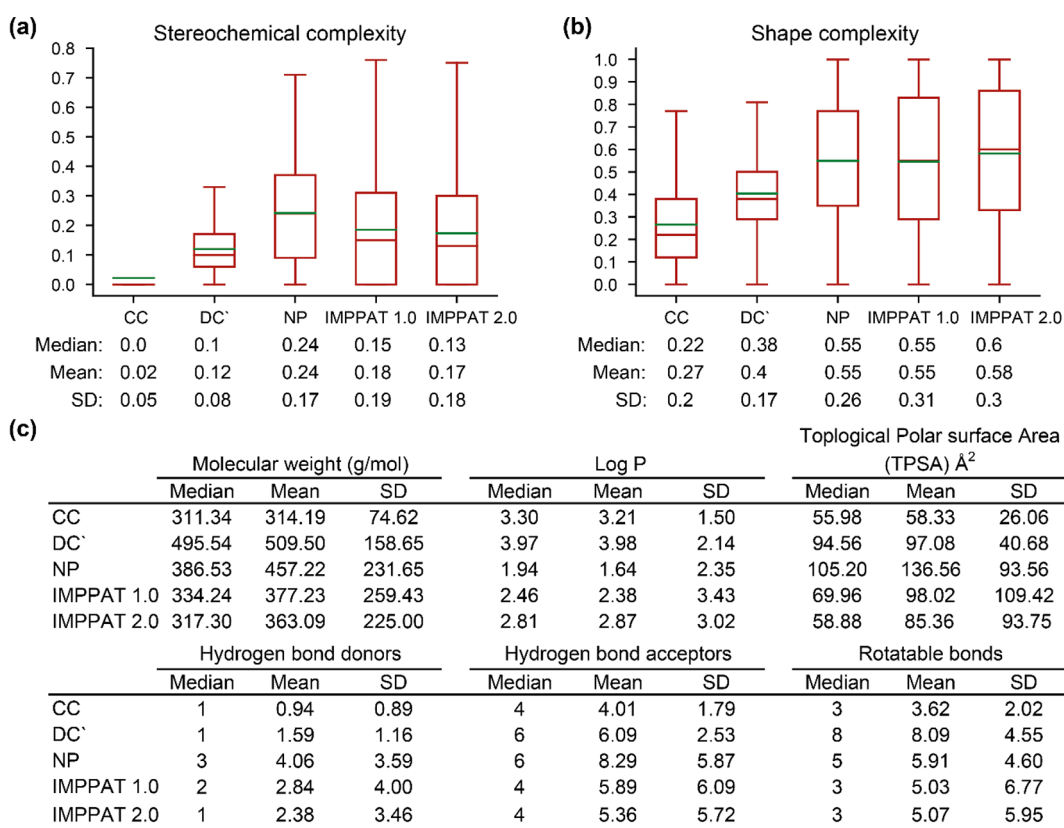
(a) Stereochemical complexity

	CC	DC`	NP	IMPPAT 1.0	IMPPAT 2.0
Median:	0.0	0.1	0.24	0.15	0.13
Mean:	0.02	0.12	0.24	0.18	0.17
SD:	0.05	0.08	0.17	0.19	0.18

(b) Shape complexity

	CC	DC`	NP	IMPPAT 1.0	IMPPAT 2.0
Median:	0.22	0.38	0.55	0.55	0.6
Mean:	0.27	0.4	0.55	0.55	0.58
SD:	0.2	0.17	0.26	0.31	0.3

(c)

	Molecular weight (g/mol)			Log P			Toplogical Polar surface Area (TPSA) Å^2		
	Median	Mean	SD	Median	Mean	SD	Median	Mean	SD
CC	311.34	314.19	74.62	3.30	3.21	1.50	55.98	58.33	26.06
DC`	495.54	509.50	158.65	3.97	3.98	2.14	94.56	97.08	40.68
NP	386.53	457.22	231.65	1.94	1.64	2.35	105.20	136.56	93.56
IMPPAT 1.0	334.24	377.23	259.43	2.46	2.38	3.43	69.96	98.02	109.42
IMPPAT 2.0	317.30	363.09	225.00	2.81	2.87	3.02	58.88	85.36	93.75

	Hydrogen bond donors			Hydrogen bond acceptors			Rotatable bonds		
	Median	Mean	SD	Median	Mean	SD	Median	Mean	SD
CC	1	0.94	0.89	4	4.01	1.79	3	3.62	2.02
DC`	1	1.59	1.16	6	6.09	2.53	8	8.09	4.55
NP	3	4.06	3.59	6	8.29	5.87	5	5.91	4.60
IMPPAT 1.0	2	2.84	4.00	4	5.89	6.09	3	5.03	6.77
IMPPAT 2.0	1	2.38	3.46	4	5.36	5.72	3	5.07	5.95

Figure 6. Comparison of the molecular complexity of chemical libraries. (a) Distribution of the stereochemical complexity and (b) the shape complexity for small molecules in five chemical libraries, namely, CC, DC', NP, IMPPAT 1.0, and IMPPAT 2.0. Note that the lower end of the box plot is the first quartile, the upper end is the third quartile, the brown line inside the box is the median, and the green line is the mean of the distribution. Also, the median, mean, and standard deviation (SD) of the distribution are shown below the box plot. (c) Median, mean, and SD for six physicochemical properties, namely, molecular weight (g/mol), log P, topological polar surface area (TPSA) (Å^2), number of hydrogen bond donors, number of hydrogen bond acceptors, and number of rotatable bonds, for small molecules in five chemical libraries.

The chemical similarity filter tab enables identification of phytochemicals in IMPPAT 2.0 that are structurally similar to a user-submitted query compound. To submit a query compound, the user can either use the molecular editor to draw its chemical structure and thereafter search the corresponding SMILES, or directly enter the SMILES to perform the search. Upon submitting the SMILES of a query compound, the webserver will display a table listing the top 10 phytochemicals in IMPPAT 2.0 that are structurally similar based on the Tanimoto coefficient (Tc),[49] a standard measure to quantify the extent of chemical similarity (Methods). The scaffold filter tab enables a user to retrieve phytochemicals based on shared molecular scaffold. A user can select one of the three types of scaffolds, namely, graph/node/bond (G/N/B) level, graph/node (G/N) level, or graph level (Methods), and thereafter select the desired scaffold from the dropdown menu to view the list of phytochemicals in the database having the desired scaffold. Overall, the advanced search page of IMPPAT 2.0 enables cheminformatics based exploration of the phytochemical space of the Indian medicinal plants toward natural product based drug discovery.

Detailed Information on Phytochemicals. In the web interface, a user is redirected to a dedicated page containing detailed information on a specific phytochemical upon clicking the corresponding phytochemical identifier or name in the tables fetched via *browse*, *basic search*, or *advanced search* options. The dedicated page provides detailed information for a phytochemical in six tabs: (a) summary, (b) physicochem-

ical, (c) drug-likeness, (d) ADMET, (e) descriptors, and (f) predicted human target proteins (Figure 5b). The summary tab provides basic information such as the chemical name, chemical classification, 2D and 3D chemical structures, and molecular scaffolds for the phytochemical. The remaining five tabs give the physicochemical properties, drug-likeness scores, predicted ADMET properties, molecular descriptors, and predicted human target proteins from the STITCH[44] database, respectively, for the phytochemical. The predicted human target proteins tab also provides a network visualization of the phytochemical−predicted human target protein associations.

Molecular Complexity Comparison with Other Collections of Small Molecules. Small molecules that are selective and specific binders of a target protein are preferable for drug development over promiscuous binders that can interact with both primary target and off-target proteins. Several molecular complexity metrics have been shown to correlate with the selectivity or promiscuity of small molecules.[50,51] In particular, Clemons *et al.*[52] have shown that stereochemical complexity and shape complexity are excellent indicators of target protein specificity of small molecules.

In their work, Clemons *et al.*[52] correlated the distribution of stereochemical and shape complexity with protein binding specificity of three different representative small molecule collections, namely, commercial compounds (CC), diversity-oriented synthesis compounds (DC'), and natural products (NP) (Methods). Clemons *et al.*[52] found that CC, DC', and

Table 2. Scaffold Diversity of Phytochemicals in IMPPAT 2.0 and Comparison with Other Chemical Libraries[a]

chemical library	M	N	N_{sing}	N/M	N_{sing}/M	N_{sing}/N	AUC	P_{50}
approved drugs	2097	1255	1012	0.6	0.48	0.81	0.69	17.93
TCM-Mesh	9417	3946	2626	0.42	0.28	0.67	0.75	11.02
NANPDB	4645	1762	1093	0.38	0.24	0.62	0.76	10.67
IMPPAT 2.0	15,226	5179	3338	0.34	0.22	0.64	0.79	6.58
NPATLAS	31,099	10,227	5947	0.33	0.19	0.58	0.79	8.35
COCONUT	385,926	109,024	65,963	0.28	0.17	0.61	0.82	4.82
CMAUP	43,987	11,105	6151	0.25	0.14	0.55	0.82	5.15
UNPD	215,585	44,281	22,514	0.21	0.1	0.51	0.85	3.39
SuperNatural II	308,998	62,125	30,453	0.2	0.1	0.49	0.85	3.61
PubChem	101,452,728	12,493,379	7,059,386	0.12	0.07	0.57	0.91	0.22

[a]The molecular scaffolds are computed at the graph/node/bond (G/N/B) level. Here, M is the number of molecules with scaffold, and this number is less than the library size as linear molecules with no ring system have no scaffolds. Further, N is the number of scaffolds, N_{sing} is the number of singleton scaffolds, AUC is the area under the curve, and P_{50} is the percentage of scaffolds that account for 50% of the chemical library.
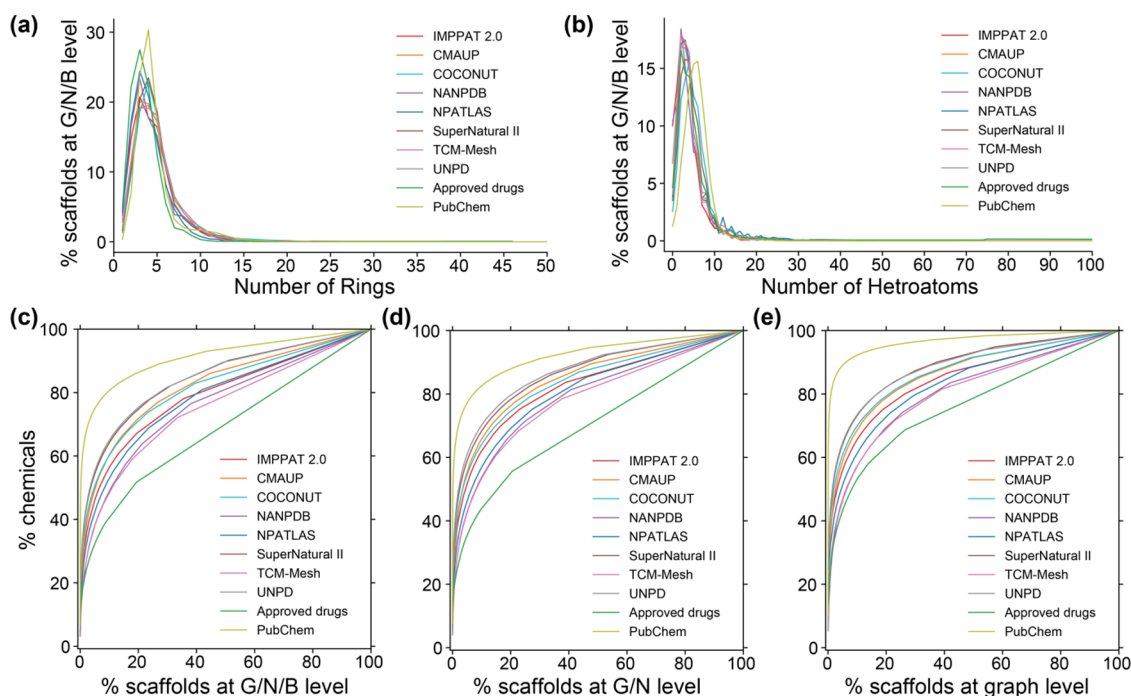

Figure 7. Analysis of the scaffold diversity of phytochemicals in IMPPAT 2.0 with seven other natural product libraries, approved drugs, and organic compounds from PubChem. Distribution of (a) the number of ring systems and (b) the number of heteroatoms in scaffolds at the graph/node/bond (G/N/B) level. Cyclic system retrieval (CSR) curves for scaffolds at the (c) G/N/B level, (d) graph/node (G/N) level, and (e) graph level.

NP molecules on average have low, intermediate, and high values, respectively, of both stereochemical and shape complexity. Thereafter, Clemons et al.[52] correlated the two molecular complexities to protein binding specificities to find that CC molecules with low complexity are enriched in promiscuous binders and depleted in specific binders, whereas, in comparison, DC' molecules with intermediate complexity and NP molecules with high complexity are more enriched in specific binders and depleted in promiscuous binders. Lastly, NP molecules were found to be more depleted in promiscuous binders in comparison to DC' molecules.[52]

Previously,[16] we compared the stereochemical and shape complexity of the CC, DC', and NP molecules with 9596 phytochemicals in IMPPAT 1.0 from Indian medicinal plants and 10,140 phytochemicals in TCM-Mesh[26] from Chinese medicinal plants. In a nutshell, we showed conclusively that phytochemicals in both IMPPAT 1.0 and TCM-Mesh are

similar to NP collection in terms of their distributions of stereochemical and shape complexity. Because of the significant increase in the number of phytochemicals in IMPPAT 2.0, we compared the distribution of stereochemical and shape complexity of CC, DC', and NP molecules with phytochemicals in IMPPAT 1.0 and IMPPAT 2.0 (Figure 6a). We find that the distributions of stereochemical and shape complexity for phytochemicals in IMPPAT 2.0 are very similar to IMPPAT 1.0 and closer to NP rather than DC' or CC collections (Figure 6a).

In another study, Clemons et al.[53] have shown that CC, DC', and NP occupy different regions in the physicochemical space defined by six properties, namely, molecular weight, log P, topological polar surface area, number of hydrogen bond donors, number of hydrogen bond acceptors, and number of rotatable bonds. In terms of these six physicochemical properties, we also find that phytochemicals in IMPPAT 2.0
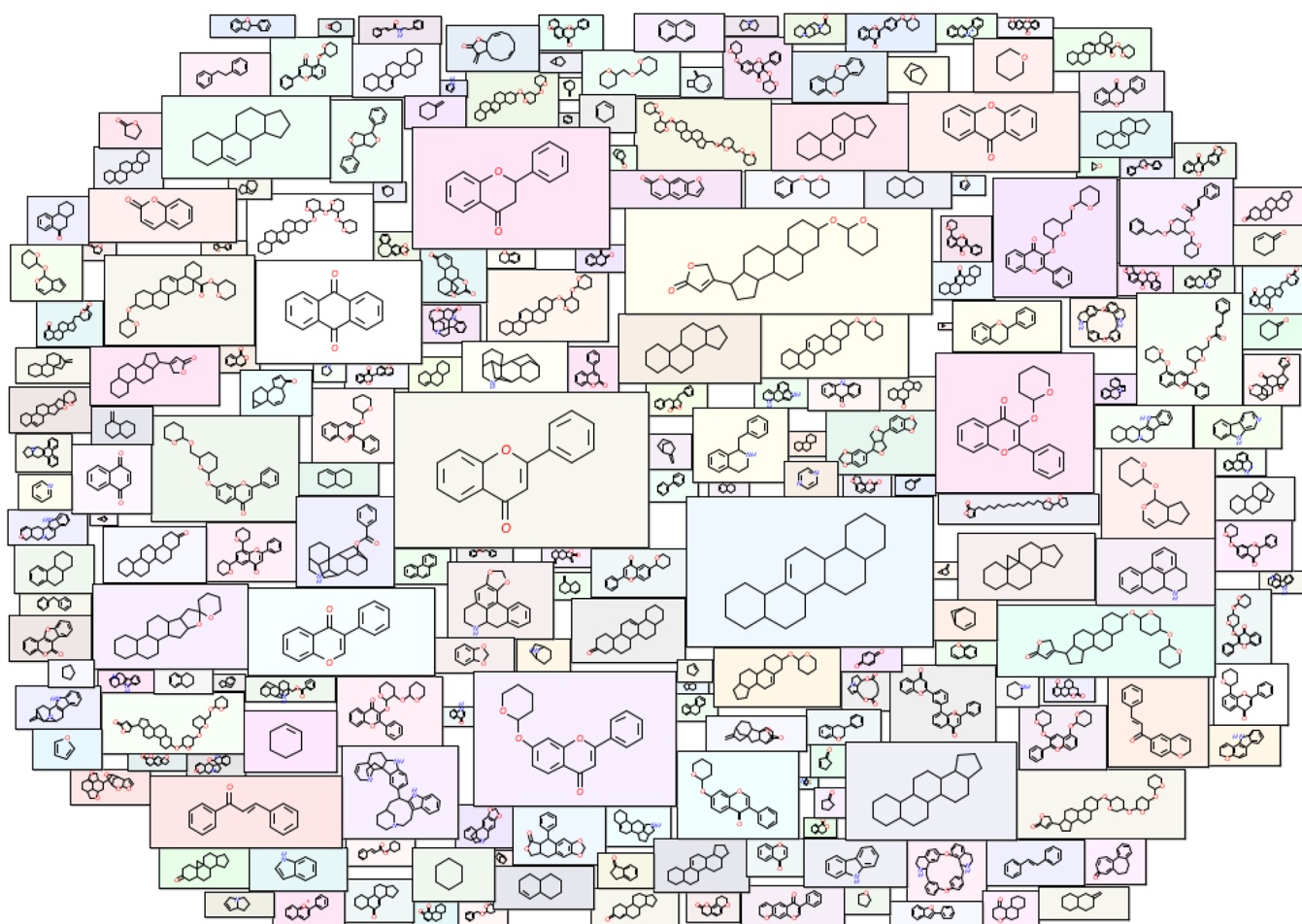
Figure 8. Molecular cloud visualization[65,66] of the top scaffolds at the G/N/B level present in phytochemicals of IMPPAT 2.0. The top comprises the 217 scaffolds at the G/N/B level that are present in ≥10 phytochemicals in IMPPAT 2.0. In this figure, 216 of these top scaffolds are shown after excluding the benzene ring (which is the most frequent scaffold in all large chemical libraries). Here, the size of the structure is proportional to the frequency of occurrence of the scaffold in phytochemicals of IMPPAT 2.0.

are very similar to IMPPAT 1.0 and closer to NP and DC' rather than CC collection (Figure 6b).

Overall, our analysis of the molecular complexities of the phytochemicals in IMPPAT 2.0 finds that the phytochemical space of Indian medicinal plants has many similarities with other natural product spaces. Notably, the phytochemical space is likely to be enriched in specific protein binders and, therefore, is a valuable space for ongoing efforts in drug discovery.

Molecular Scaffold Based Structural Diversity. Analysis of the structural diversity of a chemical space has significance for the discovery of new and novel small molecule entities. The concept of molecular scaffolds has emerged as one of the reliable ways to quantify the structural diversity[54] of chemical libraries. One way to define the molecular scaffold is via the core structure of a molecule with all its ring systems and all chain fragments connecting the rings.[54,55] Previously, Lipkus et al.[34,35] have analyzed the scaffold diversity of organic compounds compiled in the Chemical Abstracts Service (CAS) database to find that the frequency distribution of scaffolds is uneven, with most scaffolds occurring in a small number of molecules and few scaffolds occurring in a very large number of molecules. To quantify the scaffold diversity of the phytochemicals in IMPPAT 2.0, we followed Lipkus et al.[34,35] to compute the molecular scaffold at three levels,

namely, graph/node/bond (G/N/B) level, graph/node (G/N) level, and graph level (Methods). Among the phytochemicals in IMPPAT 2.0, we find 5179 scaffolds at the G/N/B level, 4072 at the G/N level, and 3434 at the graph level.

Thereafter, we compared the scaffold diversity of IMPPAT 2.0 with seven other natural product libraries (CMAUP,[56] COCONUT,[57] NANPDB,[58] NPATLAS,[59] SuperNatural-II,[60] TCM-Mesh,[26] and UNPD), approved drugs obtained from Drugbank,[61] and more than 100 million organic compounds from PubChem[33] (Table 2). Focusing solely on scaffolds at the G/N/B level, we find that phytochemical space of IMPPAT 2.0 is the third highest among the seven natural product libraries in terms of the fraction of scaffolds per molecule (N/M) and the fraction of singleton scaffolds per molecule (N_{sing}/M) after TCM-Mesh and NANPDB (Table 2).

Figure 7a,b shows the distribution of the number of rings and number of heteroatoms across the 5179 scaffolds at the G/N/B level found in phytochemicals of IMPPAT 2.0. Whereas more than 74% of the 5179 scaffolds are relatively small with ≤5 rings in them, only 2.5% of the scaffolds have ≥10 rings (Figure 7a). Notably, 231 scaffolds (4.5%) are single ring systems, and this indicates a high degree of ring diversity in phytochemicals of IMPPAT 2.0. We also find that 49.7% of the 5179 scaffolds have two, three, or four heteroatoms and that only 0.4% of the scaffolds contain ≥20 heteroatoms (Figure
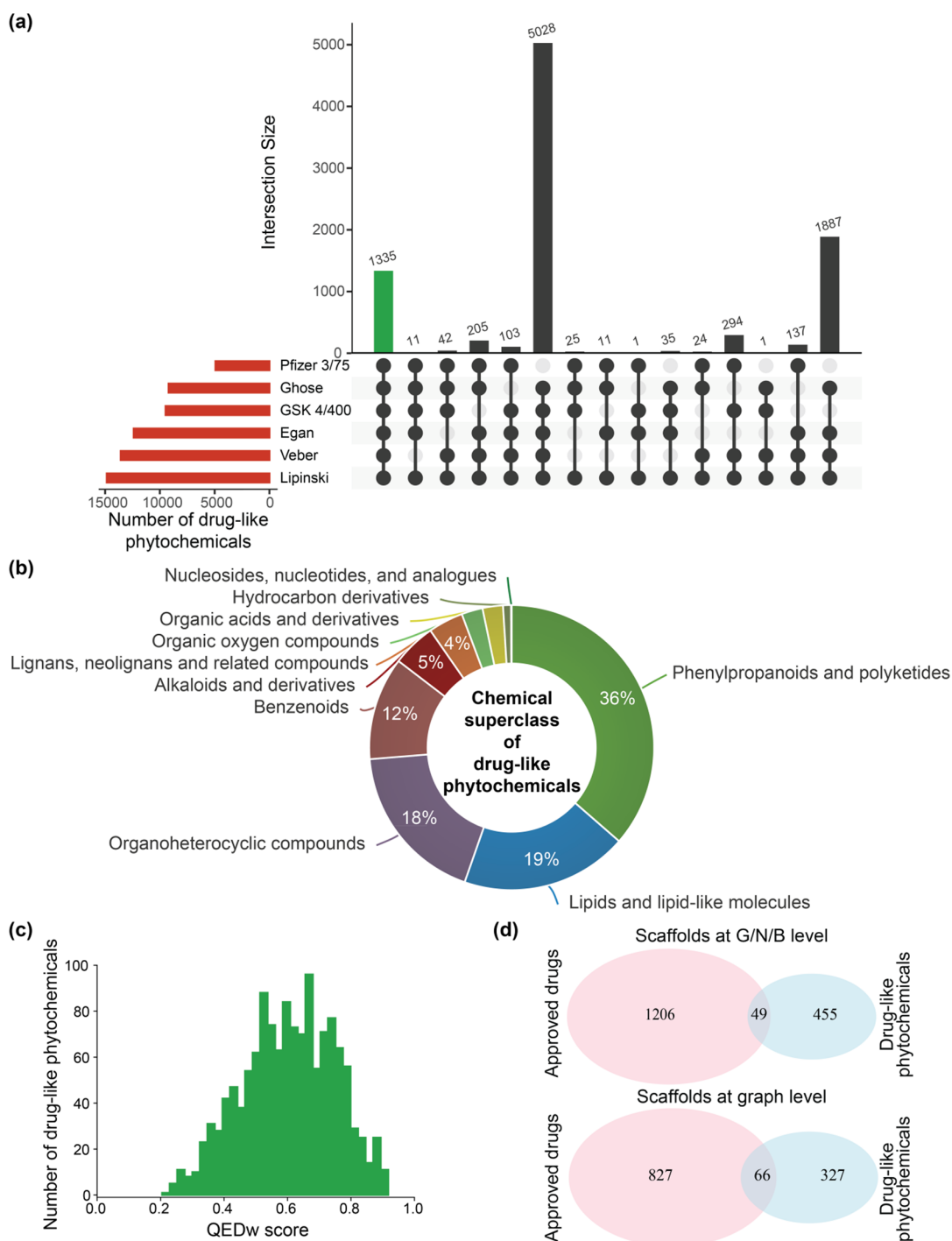
Figure 9. Drug-likeness analysis of phytochemicals in IMPPAT 2.0. (a) UpSet plot visualization of the set intersections of phytochemicals that pass one or more of the six drug-likeness rules. The horizontal bars show the number of phytochemicals that pass the different drug-likeness rules. The vertical bars show the set intersections between phytochemicals that pass different drug-likeness rules. The green bar shows the 1335 phytochemicals that pass all six drug-likeness rules. This plot was generated using the UpSetR package.[74] (b) Chemical superclass of the 1335 drug-like phytochemicals as predicted by ClassyFire. (c) Distribution of QEDw scores for the 1335 drug-like phytochemicals. (d) Common scaffolds at the graph/node/bond (G/N/B) level and the graph level between the space of 1335 drug-like phytochemicals and approved drugs.

7b). Further, 518 scaffolds (10%) are completely composed of carbon atoms. Figure 7a,b also shows that the distributions of number of rings and number of heteroatoms in scaffolds found in phytochemicals of IMPPAT 2.0 are similar to the respective distributions for other natural product libraries, approved drugs, and organic compounds from PubChem.

To further understand and compare the structural diversity of the phytochemical space of IMPPAT 2.0 with other chemical libraries, cyclic system retrieval (CSR) curves[34,35,62,63] were plotted for scaffolds computed at the G/N/B level (Figure 7c), G/N level (Figure 7d), and graph level (Figure 7e). CSR curves were generated by plotting the percent of scaffolds on the x axis and the percent of
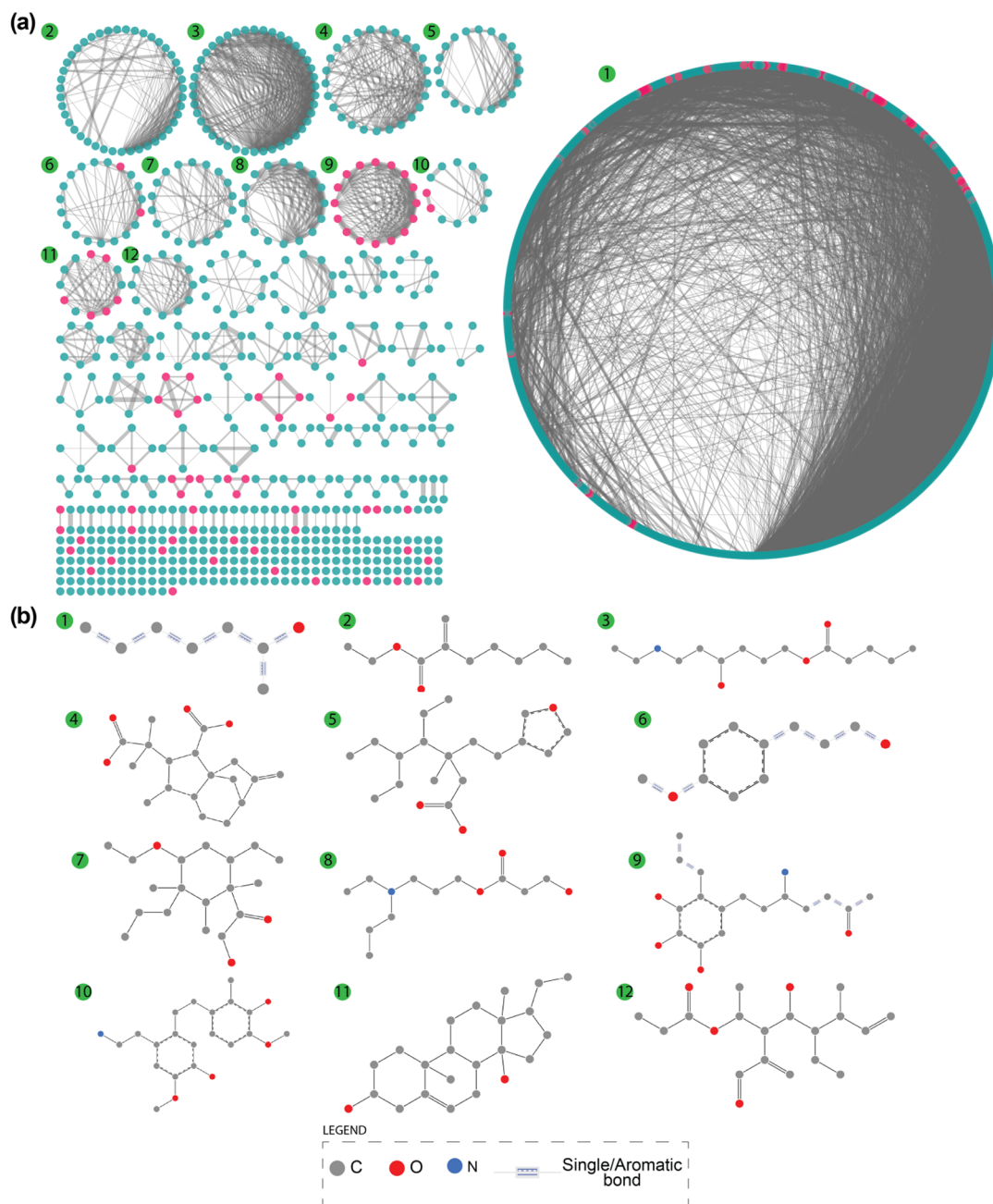
Figure 10. (a) Chemical similarity network (CSN) of the 1335 drug-like phytochemicals in IMPPAT 2.0. The degree sorted circle layout in Cytoscape[76] is used to visualize the CSN. Cyan nodes correspond to drug-like phytochemicals that are not similar to any approved drug, and pink nodes correspond to those that are similar to at least one approved drug. Edge thickness is proportional to the chemical similarity between the pair of drug-like phytochemicals. (b) Visualization of the SMARTS corresponding to the maximum common substructure (MCS) for the top 12 connected components obtained using the SMARTSview webserver.[77,78]

compounds that contain those scaffolds on the y axis. From the CSR curves, metrics such as area under the curve (AUC) and percent scaffolds required to retrieve 50% of the compounds (P_{50}) were computed. Notably, several studies have used the above metrics to quantify and compare scaffold diversity of chemical libraries.[34,35,62−64] In an ideal distribution with maximum scaffold diversity wherein each compound has a unique scaffold, the CSR curve will be the diagonal line with an AUC value of 0.5. It is seen that the CSR curves for phytochemicals in IMPPAT 2.0 (red) and other chemical libraries rise steeply and then level off (Figure 7c−e). As we move from scaffolds at the G/N/B level (least abstraction) to

the G/N level to the graph level (high abstraction), the scaffold diversity reduces across all the chemical libraries, with CSR curves shifting up away from the diagonal (Figure 7c−e).

Importantly, the scaffold diversity of phytochemicals in IMPPAT 2.0 (red) and other natural product libraries lies in between the scaffold diversity of 100 million organic compounds from PubChem (low diversity) and approved drugs (high diversity) (Figure 7c−e). Table 2 lists the AUC and P_{50} from CSR curves of scaffolds at the G/N/B level for the phytochemicals in IMPPAT 2.0 and other chemical libraries. In line with expectation, the approved drug library was found to be the most diverse with AUC of 0.69 and P_{50} of
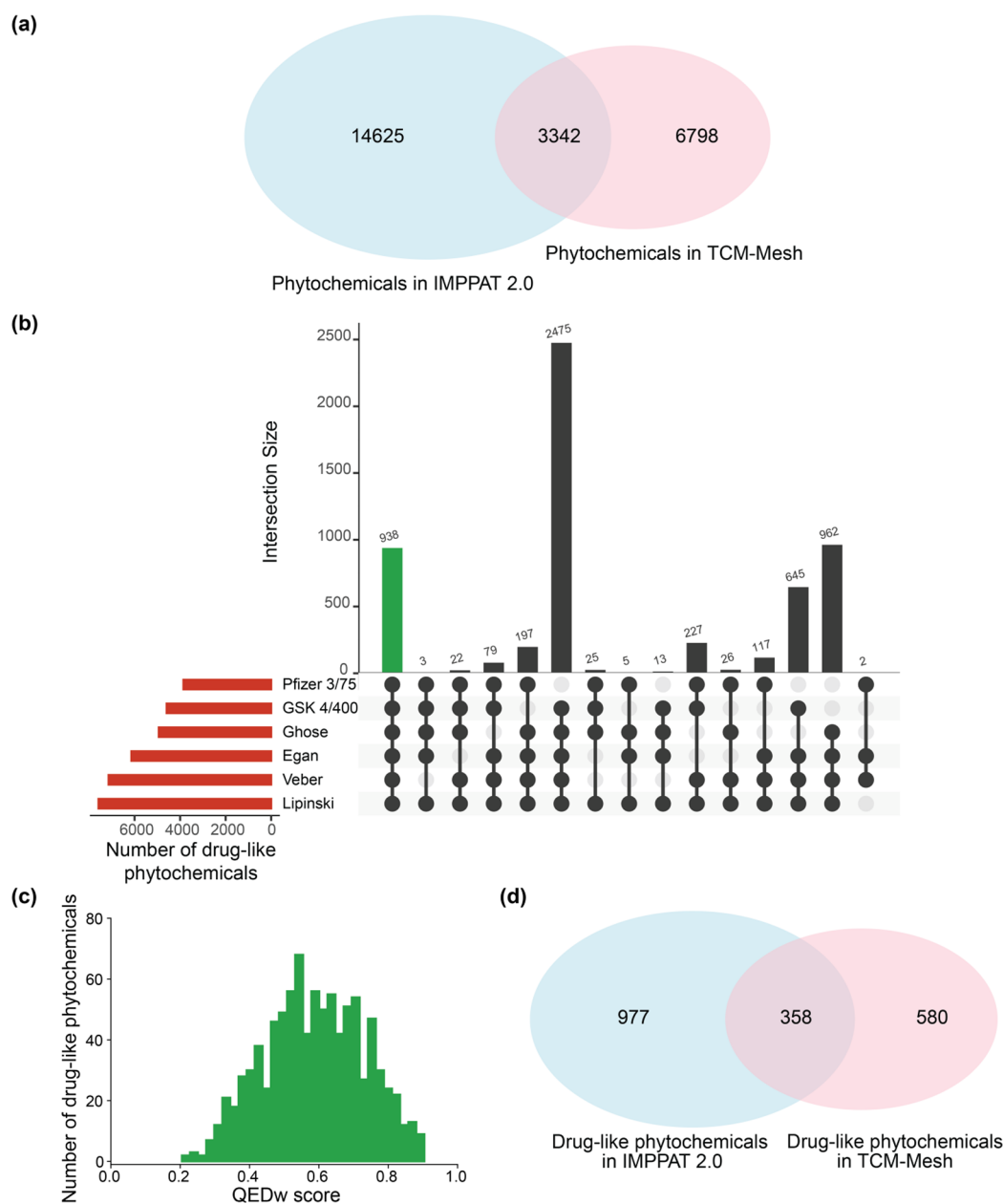
Figure 11. Comparison of the phytochemical space of Indian medicinal plants and Chinese medicinal plants. (a) Venn diagram shows the overlap between the phytochemicals in IMPPAT 2.0 and TCM-Mesh. (b) UpSet plot visualization of the set intersections of phytochemicals in TCM-Mesh that pass one or more of the six drug-likeness rules. The horizontal bars show the number of phytochemicals that pass the different drug-likeness rules. The vertical bars show the set intersections between phytochemicals that pass different drug-likeness rules. The green bar shows the 938 phytochemicals that pass all six drug-likeness rules. (c) Distribution of QEDw scores for the 938 drug-like phytochemicals in TCM-Mesh. (d) Venn diagram shows the overlap between the drug-like phytochemicals in IMPPAT 2.0 and TCM-Mesh.

17.93% (Table 2). Interestingly, the scaffold diversity of phytochemicals in IMPPAT 2.0 was found to be greater than that of the entire organic compound library from PubChem, and moreover, it is the third or fourth most diverse library among the eight natural product libraries based on AUC of 0.79 and P_{50} of 6.58% (Table 2). Further, 64.5% of the 5179 scaffolds at the G/N/B level found in phytochemicals of IMPPAT 2.0 are singletons that are present in only one compound (Table 2). In contrast, 217 scaffolds present in 10 or more phytochemicals cumulatively account for 43.6% of the phytochemicals in IMPPAT 2.0, and a molecular cloud visualization[65,66] of these scaffolds is shown in Figure 8 (after excluding the benzene ring scaffold). In sum, these

results highlight that the phytochemical space of IMPPAT 2.0 is structurally diverse with high scaffold diversity in comparison with the organic compounds from PubChem and, moreover, has similar scaffold diversity as other large natural product libraries.

Drug-like Phytochemical Space. Natural products have been an important source of approved drugs.[8,67] To predict the subset of drug-like phytochemicals in IMPPAT 2.0, we used six scoring schemes, namely, Lipinski's rule of five (RO5),[68] Ghose rule,[69] Veber rule,[70] Egan rule,[71] Pfizer 3/75 rule,[72] and GlaxoSmithKline's (GSK) 4/400 rule.[73] Figure 9a is an UpSet[74] visualization of the set intersections of phytochemicals that pass one or more of these six rules. Majority of the

phytochemicals pass RO5 (14847) followed by Veber (13574) and Egan (12390) rules. Pfizer 3/75 was found to be the most restrictive rule, with 4924 phytochemicals passing it. A drug-like subset of 1335 phytochemicals is identified based on the stringent criteria of passing all six rules (Figure 9a, Table S3).

The top five plants in IMPPAT 2.0 based on associated drug-like phytochemicals are *Senna obtusifolia* (22), *Artemisia annua* (21), *Ailanthus altissima* (19), *Catharanthus roseus* (19), and *Senna tora* (19). Figure 9b shows the chemical classification for the 1335 drug-like phytochemicals obtained using ClassyFire.[40] The top three chemical superclasses, namely, Phenylpropanoids and polyketides, Lipids and lipid-like molecules, and Organoheterocyclic compounds, account for 486, 253, and 245 drug-like phytochemicals, respectively.

The weighted quantitative estimate of drug-likeness (QEDw) score can also be used to assess the drug-likeness of small molecules, and this measure can take values between 0 (least drug-like) to 1 (most drug-like).[75] For the 1335 drug-like phytochemicals, Figure 9c shows the distribution of QEDw scores with a mean of 0.60 and a standard deviation of 0.14. Notably, 104 of the drug-like phytochemicals have a high QEDw score of ≥0.80.

We also compared the 1335 drug-like phytochemicals in IMPPAT 2.0 with the drugs approved by the United States Federal Drug Administration (US FDA). A set of 2567 approved drugs was obtained from DrugBank[61] version 5.1.9. On the basis of chemical similarity (Tc ≥ 0.50; Methods), we find 130 drug-like phytochemicals to be similar to one or more approved drugs. Interestingly, 11 drug-like phytochemicals in IMPPAT 2.0 are already US FDA approved drugs.

To assess the overlap in core chemical structure, we next computed the molecular scaffolds for the 1335 drug-like phytochemicals and 2567 approved drugs. At the G/N/B, G/N, and graph levels, the 1335 drug-like phytochemicals were found to have 504, 444, and 393 scaffolds, respectively, whereas the 2567 approved drugs have 1255, 1171, and 893 scaffolds, respectively. Importantly, the drug-like phytochemicals and approved drugs share only 49, 60, and 66 scaffolds at the G/N/B, G/N, and graph levels, respectively (Figure 9d). Thus, the drug-like phytochemicals in IMPPAT 2.0 present a unique chemical scaffold space with minimal overlap with approved drugs. These results highlight the potential of IMPPAT 2.0 in aiding the ongoing hunt for new bioactive molecules.

By constructing a chemical similarity network (CSN), we next analyzed the structural diversity of the drug-like space of 1335 phytochemicals (Methods). Figure 10a shows the drug-like CSN wherein nodes correspond to phytochemicals and an edge exists between any pair of phytochemicals if Tc ≥ 0.5. The drug-like CSN is very sparse with a graph density of 0.01, and it can be partitioned into 90 connected components (with at least two nodes each) and 210 isolated nodes. In Figure 10a, the top 12 connected components in terms of the number of constituent nodes are labeled. For instance, the connected component labeled 9 consists of 16 phytochemicals of which two (Colchicine and its metabolite Colchiceine) are approved drugs and the remaining phytochemicals are similar to them. For each of the top 12 components, the maximum common substructure (MCS) is shown in Figure 10b; the substructures confirm the structural uniqueness of the different connected components (Methods). In sum, the CSN highlights the chemical dissimilarity and, hence, the structural diversity of the drug-like space of 1335 phytochemicals.

Comparison with the Phytochemical Space of Chinese Medicinal Plants. Previously,[16] a comparison of the 9596 phytochemicals in IMPPAT 1.0 with the 10,140 phytochemicals in TCM-Mesh[26] revealed that less than 25% of phytochemicals (2305) in IMPPAT 1.0 are present in the TCM-Mesh. Notably, TCM-Mesh is a large-scale database compiling information on 10,140 phytochemicals produced by 6235 Chinese medicinal plants.[26] We also performed a comparison of the 17,967 phytochemicals in IMPPAT 2.0 with the 10,140 phytochemicals in TCM-Mesh. Although the number of phytochemicals common to IMPPAT 2.0 and TCM-Mesh has increased to 3342, the percentage of the phytochemical space of IMPPAT 2.0 that is shared with TCM-Mesh has decreased to 18.6% (Figure 11a).

Further, we compared the drug-like subset of 1335 phytochemicals in IMPPAT 2.0 with the corresponding drug-like subset in TCM-Mesh (Methods). Specifically, a subset of 938 drug-like phytochemicals was obtained in TCM-Mesh based on the six rules (Figure 11b). Further, Figure 11c shows the distribution of QEDw scores for the 938 drug-like phytochemicals in TCM-Mesh, and this distribution has a mean value of 0.59 and standard deviation of 0.14, similar to the distribution for the 1335 drug-like phytochemicals in IMPPAT 2.0. Lastly, there is a minor overlap of 338 phytochemicals between the subsets of drug-like phytochemicals in IMPPAT 2.0 and TCM-Mesh. These analyses attest to the uniqueness of the phytochemical spaces of Indian herbs and Chinese herbs, and therefore, the phytochemical atlas IMPPAT 2.0 is expected to further enrich the space of natural products.

■ CONCLUSIONS

In this contribution, we present IMPPAT 2.0, an enhanced and expanded database compiling information via extensive manual curation on Indian medicinal plants, their phytochemicals, and their therapeutic uses. In the updated database, we have more than doubled the coverage of Indian medicinal plants and nearly doubled the size of the phytochemical space. Further, we compile the phytochemicals and therapeutic uses of the Indian medicinal plants at the level of plant parts. At the level of associations, IMPPAT 2.0 compiles 189,386 plant–part–phytochemical associations and 89,733 plant–part–therapeutic use associations. Importantly, IMPPAT 2.0 provides a FAIR[24]-compliant nonredundant *in silico* stereo-aware library of 17,967 phytochemicals. The phytochemical library has been annotated with several features including 2D and 3D chemical structures, molecular scaffolds, predicted human target proteins, physicochemical properties, drug-likeness scores, and predicted ADMET properties. This will enable the effective use of the phytochemical library for screening efforts toward drug discovery. Also, the 1095 standardized therapeutic use terms in IMPPAT 2.0 are mapped to standard medical terms used in western medicine. The IMPPAT 2.0 web interface has been completely redesigned to facilitate ease of use and to serve as a cheminformatics platform for exploring the phytochemical space of Indian medicinal plants.

The cheminformatics analysis of the phytochemicals in IMPPAT 2.0 revealed that their stereochemical complexity and shape complexity are similar to those of the other natural products. Our analysis suggests that, like the library in IMPPAT 1.0, the phytochemicals in IMPPAT 2.0 are also more likely to be enriched with specific protein binders rather than promiscuous binders. The structural diversity analysis

using molecular scaffolds has shown that the phytochemicals in IMPPAT 2.0 are structurally diverse with scaffold diversity similar to large natural product databases. We identified 1335 phytochemicals in IMPPAT 2.0 as drug-like, of which only 11 phytochemicals were identified as approved drugs. Further, the chemical similarity network of the drug-like phytochemicals highlights the structural diversity of the drug-like space in IMPPAT 2.0. Finally, the comparison with the phytochemicals from Chinese medicinal plants shows that there is minimal overlap with the phytochemicals from Indian medicinal plants compiled in IMPPAT 2.0. These results show the uniqueness of the phytochemical space of IMPPAT 2.0 and its potential to further enrich the natural product chemical space.

In the future, we will continue to expand, enhance, and develop this unique platform to explore the phytochemical space of Indian medicinal plants. Also, the collection and standardization of the information on traditional Indian medicinal formulations from pharmacopeias and several books in vernacular languages remain a challenge. Further, unclear policies on access and reuse of data on traditional Indian medicinal formulations in other databases have dissuaded us from linking to such data in IMPPAT 2.0. However, we hope to incorporate manually curated information on traditional Indian medicinal formulations and the Indian medicinal plants used in them in the next update of IMPPAT. In conclusion, IMPPAT 2.0 is a unique database enabling computational and experimental research in the area of natural product and traditional knowledge based drug discovery.

■ METHODS

Plant Annotation. For the 4010 Indian medicinal plants in IMPPAT 2.0, the taxonomic information on kingdom, family, and group was compiled using The Plant List database.[28] The common names of the Indian medicinal plants were obtained from the Flowers of India database,[79] which compiles information for more than 6000 Indian plants. The IUCN Red List of Threatened species[29] is the most comprehensive resource on the global conservation status of animals, fungi, and plant species, and this list was used to ascertain the extinction risk of Indian medicinal plants. The usage of Indian medicinal plants in different traditional Indian systems of medicine such as Ayurveda, Siddha, Unani, Sowa-Rigpa, and Homeopathy was manually compiled from pharmacopeias published by the Government of India.

For the Indian medicinal plants in IMPPAT 2.0, we provide cross-reference links to associated information in other standard databases such as The Plant List,[28] Tropicos,[80] Encyclopedia of Indian medicinal plants from FRLHT,[81] Medicinal Plant Names Services (MPNS),[82] International Plant Names Index (IPNI),[83] Plants of the World Online (POW),[84] World Flora Online (WFO),[85] and Gardeners' World.[86]

Phytochemical Information. The 2D chemical structures of phytochemicals were converted to SDF, MOL, and MOL2 file formats using OpenBabel.[87] The images of the 2D structures of phytochemicals were generated using RDKit.[88] The 3D chemical structures of phytochemicals were retrieved from PubChem.[33] If the 3D structure for a phytochemical was not available in PubChem, the 3D structure was generated from its 2D structure using RDKit by first embedding the 2D structure using the ETKDG method and thereafter energy minimizing the structure using the MMFF94 force field.[88] The

3D structures of phytochemicals were converted to SDF, MOL, MOL2, PDB, and PDBQT file formats using Open-Babel.[87] Note that IMPPAT 2.0 provides 3D structures for 17,910 phytochemicals as the generation of 3D structures failed for the remaining 57 phytochemicals in the database. Lastly, the chemical structure of each phytochemical in SMILES, InChI, and InChIKey formats was also generated using OpenBabel.[87]

Using ClassyFire,[40] the chemical classification for each phytochemical into hierarchical levels, namely, kingdom, superclass, class, and subclass, was predicted. Further, using NP classifier,[38] a natural product specific chemical classification for each phytochemical into the biosynthetic pathway, superclass and class were predicted. For each phytochemical in our database, external links to other standard chemical databases are provided using UniChem.[89] Lastly, the natural product likeness or NP-likeness score for each phytochemical was computed using a custom RDKit script.[39,41]

For each phytochemical in our database, the physicochemical properties and drug-likeness scores were computed using in-house custom RDKit scripts. Further, the absorption, distribution, metabolism, excretion, and toxicity (ADMET) properties of the phytochemicals were predicted using SwissADME.[90] Because the SwissADME restricts the input molecules based on their length of SMILES, ADMET predictions could not be obtained for 493 phytochemicals in our database. Finally, we computed 1875 molecular descriptors, both 2D and 3D descriptors, for each phytochemical in our database using the PaDEL[91] software.

The predicted human target proteins of phytochemicals were obtained from the STITCH database.[44] Only high confidence phytochemical–human target protein interactions with a score of at least 700 were retrieved from the STITCH database. Further, the genes corresponding to the target human proteins were mapped to the HUGO Gene Nomenclature Committee (HGNC) symbols and identifiers.[92]

Phytochemicals with published experimental evidence of acting as covalent inhibitors were identified and compiled from CovalentInDB[93] and CovPDB[94] via a comparison of the chemical structures followed by manual verification.

Molecular Complexity. The molecular complexity of the phytochemicals in IMPPAT 2.0 was compared with four chemical spaces, namely, phytochemicals in IMPPAT 1.0 and three collections of small molecules obtained from Clemons et al.[52] corresponding to 6152 commercial compounds (CC), 5963 diversity-oriented synthesis compounds (DC'), and 2477 natural products (NP). For each compound in the above-mentioned five chemical spaces, we computed using RDKit[88] two size-independent metrics, namely, stereochemical complexity, which is the fraction of stereogenic carbon atoms in a compound, and shape complexity, which is the ratio of sp^3-hybridized carbon atoms to the total number of sp^2- and sp^3-hybridized carbon atoms in a compound, and six other physicochemical properties, namely, molecular weight, log P, topological polar surface area, number of hydrogen bond donors, number of hydrogen bond acceptors, and number of rotatable bonds.

Molecular Scaffold. Based on the definition by Lipkus et al.,[34,35] molecular scaffolds were computed at three levels, namely, graph/node/bond (G/N/B) level, graph/node (G/N) level, and graph level, using RDKit.[95] Scaffolds were computed by modifying the MurckoScaffold.py from RDKit.[95] The scaffold at the G/N/B level has connectivity, element, and

bond information; that at the G/N level has connectivity and element information but ignores bond information; and that at the graph level only has connectivity information.[34,35]

Quantifying and Visualizing Chemical Similarity. Chemical structure similarity between any two molecules is quantified using the widely used metric, Tanimoto coefficient (Tc),[49] which was computed using Extended Circular Fingerprints (ECFP4) as implemented in RDKit.[95] A chemical similarity network (CSN) consists of nodes corresponding to phytochemicals and edges connecting pairs of nodes with Tc ≥ 0.5. The value of Tc for a pair of molecules in the CSN gives the extent of chemical similarity between them, and this is captured by the thickness of the corresponding edge (Figure 10a). The maximum common substructure (MCS) for phytochemicals in a connected component of the CSN was computed using the FindMCS function in RDKit.[95] The SMARTS for an MCS was visualized using the SMARTSview webserver.[77,78]

Web Interface and Database Management. The IMPPAT 2.0 database has a user-friendly web interface and can be accessed at https://cb.imsc.res.in/imppat. The website is also mirrored at https://www.imppat.com/ and https://www.imppat.in/. The website is hosted on a local Apache server running on a Debian 9.1.3 Linux operating system. The association tables are stored in SQL format created using the open-source relational database management system MariaDB. The front-end of the website was created using the open-source CSS framework Bootstrap 4.1.3 customized with in-house HTML, PHP, CSS, JavaScript, and jQuery scripts. Further, Cytoscape.js[96] and jQuery plug-in DataTables are incorporated for visualizing networks and for displaying tables, respectively. Also, JSME Molecule Editor[97] and JSmol[98] are incorporated to enable drawing of chemical structures and to visualize 3D chemical structures, respectively.

Data Availability. The IMPPAT 2.0 database on phytochemicals of Indian medicinal plants is accessible via the associated website: https://cb.imsc.res.in/imppat. The compiled information in IMPPAT 2.0 is made available under a Creative Commons Attribution-NonCommercial 4.0 (CC BY-NC 4.0) International License. The computer codes used to analyze the phytochemical space of IMPPAT 2.0 are available via the associated GitHub repository: https://github.com/asamallab/imppat2.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information
The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.3c00156.

> The list of 70 books from which plant−part− phytochemical associations for Indian medicinal plants in IMPPAT 2.0 were obtained (Table S1); the list of 146 books from which plant−part−therapeutic use associations for Indian medicinal plants in IMPPAT 2.0 were obtained (Table S2); and the IMPPAT Phytochemical identifier, Chemical name, SMILES, InChI, and QEDw score for the 1335 drug-like phytochemicals in IMPPAT 2.0 identified in this study (Table S3) (XLSX)

■ AUTHOR INFORMATION

Corresponding Author
Areejit Samal − *The Institute of Mathematical Sciences (IMSc), Chennai 600113, India; Homi Bhabha National Institute (HBNI), Mumbai 400094, India;* ⓘ orcid.org/0000-0002-6796-9604; Email: asamal@imsc.res.in

Authors
R.P. Vivek-Ananth − *The Institute of Mathematical Sciences (IMSc), Chennai 600113, India; Homi Bhabha National Institute (HBNI), Mumbai 400094, India;* ⓘ orcid.org/0000-0002-3232-3299

Karthikeyan Mohanraj − *The Institute of Mathematical Sciences (IMSc), Chennai 600113, India;* Present Address: Institute for Clinical Chemistry and Laboratory Medicine, Technische Universität Dresden, Dresden 01307, Germany; ⓘ orcid.org/0000-0003-2792-0627

Ajaya Kumar Sahoo − *The Institute of Mathematical Sciences (IMSc), Chennai 600113, India; Homi Bhabha National Institute (HBNI), Mumbai 400094, India;* ⓘ orcid.org/0000-0003-3543-8021

Complete contact information is available at:
https://pubs.acs.org/10.1021/acsomega.3c00156

Author Contributions
R.P.V., K.M., and A.S. designed research. R.P.V., K.M., and A.K.S carried out the data compilation and curation. R.P.V., K.M., and A.K.S. designed the database platform and visual interface. R.P.V. performed the computational analysis. A.S. and R.P.V. wrote the manuscript. A.S. conceived and supervised the project. All authors have read and approved the manuscript.

Notes
The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

■ REFERENCES

(1) Gurib-Fakim, A. Medicinal Plants: Traditions of Yesterday and Drugs of Tomorrow. *Mol. Aspects Med.* **2006**, *27*, 1−93.

(2) Petrovska, B. Historical Review of Medicinal Plants′ Usage. *Pharmacogn. Rev.* **2012**, *6*, 1−5.

(3) Patwardhan, B. Ethnopharmacology and Drug Discovery. *J. Ethnopharmacol.* **2005**, *100*, 50−52.

(4) Patwardhan, B.; Mashelkar, R. A. Traditional Medicine-Inspired Approaches to Drug Discovery: Can Ayurveda Show the Way Forward? *Drug Discovery Today* **2009**, *14*, 804−811.

(5) Mukherjee, P. K.; Rai, S.; Kumar, V.; Mukherjee, K.; Hylands, P.; Hider, R. Plants of Indian Origin in Drug Discovery. *Expert Opin. Drug Discovery* **2007**, *2*, 633−657.

(6) Ahmad, S.; Zahiruddin, S.; Parveen, B.; Basist, P.; Parveen, A.; Gaurav; Parveen, R.; Ahmad, M. Indian Medicinal Plants and Formulations and Their Potential Against COVID-19—Preclinical and Clinical Research. *Front. Pharmacol.* **2021**, *11*, No. 578970.

(7) Grigalunas, M.; Brakmann, S.; Waldmann, H. Chemical Evolution of Natural Product Structure. *J. Am. Chem. Soc.* **2022**, *144*, 3314−3329.

(8) Newman, D. J.; Cragg, G. M. Natural Products as Sources of New Drugs over the Nearly Four Decades from 01/1981 to 09/2019. *J. Nat. Prod.* **2020**, *83*, 770−803.

(9) Atanasov, A. G. O.; Zotchev, S. B.; Dirsch, V. M.; International Natural Product Sciences Taskforce; Supuran, C. T. Natural Products in Drug Discovery: Advances and Opportunities. *Nat. Rev. Drug Discovery* **2021**, *20*, 200−216.

(10) Thomford, N.; Senthebane, D. A.; Rowe, A.; Munro, D.; Seele, P.; Maroyi, A.; Dzobo, K. Natural Products for Drug Discovery in the 21st Century: Innovations for Novel Drug Discovery. *Int. J. Mol. Sci.* **2018**, *19*, 1578.

(11) Gaston, T. E.; Mendrick, D. L.; Paine, M. F.; Roe, A. L.; Yeung, C. K. "Natural" Is Not Synonymous with "Safe:" Toxicity of Natural Products Alone and in Combination with Pharmaceutical Agents. *Regul. Toxicol. Pharmacol.* **2020**, *113*, No. 104642.

(12) Mukherjee, P. K.; Wahile, A. Integrated Approaches towards Drug Development from Ayurveda and Other Indian System of Medicines. *J. Ethnopharmacol.* **2006**, *103*, 25−35.

(13) Kirchmair, J. Molecular Informatics in Natural Products Research. *Mol. Inf.* **2020**, *39*, 2000206.

(14) Chen, Y.; Kirchmair, J. Cheminformatics in Natural Product-Based Drug Discovery. *Mol. Inf.* **2020**, *39*, 2000171.

(15) Saldívar-González, F. I.; Aldas-Bulos, V. D.; Medina-Franco, J. L.; Plisson, F. Natural Product Drug Discovery in the Artificial Intelligence Era. *Chem. Sci.* **2022**, *13*, 1526−1546.

(16) Mohanraj, K.; Karthikeyan, B. S.; Vivek-Ananth, R. P.; Chand, R. P. B.; Aparna, S. R.; Mangalapandi, P.; Samal, A. IMPPAT: A Curated Database of Indian Medicinal Plants, Phytochemistry And Therapeutics. *Sci. Rep.* **2018**, *8*, 4329.

(17) Sorokina, M.; Steinbeck, C. Review on Natural Products Databases: Where to Find Data in 2020. *J. Chem.* **2020**, *12*, 20.

(18) Vivek-Ananth, R. P.; Rana, A.; Rajan, N.; Biswal, H. S.; Samal, A. In Silico Identification of Potential Natural Product Inhibitors of Human Proteases Key to SARS-CoV-2 Infection. *Molecules* **2020**, *25*, 3822.

(19) Vivek-Ananth, R. P.; Krishnaswamy, S.; Samal, A. Potential Phytochemical Inhibitors of SARS-CoV-2 Helicase Nsp13: A Molecular Docking and Dynamic Simulation Study. *Mol. Diversity* **2022**, *26*, 429−442.

(20) Vivek-Ananth, R. P.; Sahoo, A. K.; Srivastava, A.; Samal, A. Virtual Screening of Phytochemicals from Indian Medicinal Plants against the Endonuclease Domain of SFTS Virus L Polymerase. *RSC Adv.* **2022**, *12*, 6234−6247.

(21) Basu, A.; Sarkar, A.; Maulik, U. Molecular Docking Study of Potential Phytochemicals and Their Effects on the Complex of SARS-CoV2 Spike Protein and Human ACE2. *Sci. Rep.* **2020**, *10*, 17699.

(22) Prasanth, D. S. N. B. K.; Murahari, M.; Chandramohan, V.; Panda, S. P.; Atmakuri, L. R.; Guntupalli, C. *In Silico* Identification of Potential Inhibitors from *Cinnamon* against Main Protease and Spike Glycoprotein of SARS CoV-2. *J. Biomol. Struct. Dyn.* **2021**, *39*, 4618−4632.

(23) Borkotoky, S.; Banerjee, M. A Computational Prediction of SARS-CoV-2 Structural Protein Inhibitors from *Azadirachta Indica* (Neem). *J. Biomol. Struct. Dyn.* **2021**, *39*, 4111−4121.

(24) Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; Bouwman, J.; Brookes, A. J.; Clark, T.; Crosas, M.; Dillo, I.; Dumon, O.; Edmunds, S.; Evelo, C. T.; Finkers, R.; Gonzalez-Beltran, A.; Gray, A. J. G.; Groth, P.; Goble, C.; Grethe, J. S.; Heringa, J.; 't Hoen, P. A. C.; Hooft, R.; Kuhn, T.; Kok, R.; Kok, J.; Lusher, S. J.; Martone, M. E.; Mons, A.; Packer, A. L.; Persson, B.; Rocca-Serra, P.; Roos, M.; van Schaik, R.; Sansone, S.-A.; Schultes, E.;

Sengstag, T.; Slater, T.; Strawn, G.; Swertz, M. A.; Thompson, M.; van der Lei, J.; van Mulligen, E.; Velterop, J.; Waagmeester, A.; Wittenburg, P.; Wolstencroft, K.; Zhao, J.; Mons, B. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci. Data* **2016**, *3*, 160018.

(25) Ntie-Kang, F.; Zofou, D.; Babiaka, S. B.; Meudom, R.; Scharfe, M.; Lifongo, L. L.; Mbah, J. A.; Mbaze, L. M.; Sippl, W.; Efange, S. M. N. AfroDb: A Select Highly Potent and Diverse Natural Product Library from African Medicinal Plants. *PLoS One* **2013**, *8*, No. e78085.

(26) Zhang, R.; Yu, S.; Bai, H.; Ning, K. TCM-Mesh: The Database and Analytical System for Network Pharmacology Analysis for TCM Preparations. *Sci. Rep.* **2017**, *7*, 2821.

(27) Gómez-García, A.; Medina-Franco, J. L. Progress and Impact of Latin American Natural Product Databases. *Biomolecules* **2022**, *12*, 1202.

(28) *The Plant List (2013). Version 1.1*; The Plant List http://www.theplantlist.org/ (accessed 2021-12-01).

(29) IUCN. 2021. *The IUCN Red List of Threatened Species. Version 2021−3*; IUCN https://www.iucnredlist.org (accessed 2022-03-31).

(30) Kuete, V.; Viertel, K.; Efferth, T. Antiproliferative Potential of African Medicinal Plants. In *Medicinal Plant Research in Africa*; Elsevier, 2013; pp. 711−724. DOI: 10.1016/B978-0-12-405927-6.00018-7.

(31) Kinghorn, A. D. Reviews on Indian Medicinal Plants, Vols. 1−3 (Abe-Alle; Alli-Ard; Are-Azi) Edited by A. K. Gupta and N. Tandon, Assisted by M. Sharma (Indian Council of Medical Research, New Delhi). 2004. *J. Nat. Prod.* **2005**, *68*, 153−154.

(32) Pathania, S.; Ramakrishnan, S. M.; Randhawa, V.; Bagler, G. SerpentinaDB: A Database of Plant-Derived Molecules of Rauvolfia Serpentina. *BMC Complementary Altern. Med.* **2015**, *15*, 262.

(33) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem in 2021: New Data Content and Improved Web Interfaces. *Nucleic Acids Res.* **2021**, *49*, D1388−D1395.

(34) Lipkus, A. H.; Yuan, Q.; Lucas, K. A.; Funk, S. A.; Bartelt, W. F., III; Schenck, R. J.; Trippe, A. J. Structural Diversity of Organic Chemistry. A Scaffold Analysis of the CAS Registry. *J. Org. Chem.* **2008**, *73*, 4443−4451.

(35) Lipkus, A. H.; Watkins, S. P.; Gengras, K.; McBride, M. J.; Wills, T. J. Recent Changes in the Scaffold Diversity of Organic Chemistry As Seen in the CAS Registry. *J. Org. Chem.* **2019**, *84*, 13948−13956.

(36) Ertl, P. An Algorithm to Identify Functional Groups in Organic Molecules. *J. Chem.* **2017**, *9*, 36.

(37) O'Boyle, N.; Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. 2018, DOI: 10.26434/chemrxiv.7097960.v1.

(38) Kim, H. W.; Wang, M.; Leber, C. A.; Nothias, L.-F.; Reher, R.; Kang, K. B.; van der Hooft, J. J. J.; Dorrestein, P. C.; Gerwick, W. H.; Cottrell, G. W. NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *J. Nat. Prod.* **2021**, *84*, 2795−2807.

(39) *NP-Likeness score*; 2022. https://github.com/rdkit/rdkit/tree/master/Contrib/NP_Score (accessed 2022-03-31).

(40) Djoumbou Feunang, Y.; Eisner, R.; Knox, C.; Chepelev, L.; Hastings, J.; Owen, G.; Fahy, E.; Steinbeck, C.; Subramanian, S.; Bolton, E.; Greiner, R.; Wishart, D. S. ClassyFire: Automated Chemical Classification with a Comprehensive, Computable Taxonomy. *J. Chem.* **2016**, *8*, 61.

(41) Ertl, P.; Roggo, S.; Schuffenhauer, A. Natural Product-Likeness Score and Its Application for Prioritization of Compound Libraries. *J. Chem. Inf. Model.* **2008**, *48*, 68−74.

(42) Sorokina, M.; Steinbeck, C. NaPLeS: A Natural Products Likeness Scorer—Web Application and Database. *J. Chem.* **2019**, *11*, 55.

(43) Vanii Jayaseelan, K.; Moreno, P.; Truszkowski, A.; Ertl, P.; Steinbeck, C. Natural Product-Likeness Score Revisited: An Open-Source, Open-Data Implementation. *BMC Bioinf.* **2012**, *13*, 106.

(44) Szklarczyk, D.; Santos, A.; von Mering, C.; Jensen, L. J.; Bork, P.; Kuhn, M. STITCH 5: Augmenting Protein-Chemical Interaction Networks with Tissue and Affinity Data. *Nucleic Acids Res.* **2016**, *44*, D380−D384.

(45) (Medical Subject Headings) ; National Library of Medicine: 2022. https://meshb.nlm.nih.gov/ (accessed 2022-03-31).

(46) *ICD-11 for Mortality and Morbidity Statistics. Version: 02/2022.* https://icd.who.int/browse11/ (accessed 2022-03-31).

(47) *Unified Medical Language System*; U.S. National Library of Medicine 2022. https://uts.nlm.nih.gov/uts/umls (accessed 2022-03-31).

(48) Schriml, L. M.; Mitraka, E.; Munro, J.; Tauber, B.; Schor, M.; Nickle, L.; Felix, V.; Jeng, L.; Bearer, C.; Lichenstein, R.; Bisordi, K.; Campion, N.; Hyman, B.; Kurland, D.; Oates, C. P.; Kibbey, S.; Sreekumar, P.; Le, C.; Giglio, M.; Greene, C. Human Disease Ontology 2018 Update: Classification, Content and Workflow Expansion. *Nucleic Acids Res.* **2019**, *47*, D955−D962.

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

(50) Méndez-Lucio, O.; Medina-Franco, J. L. The Many Roles of Molecular Complexity in Drug Discovery. *Drug Discovery Today* **2017**, *22*, 120−126.

(51) Vivek-Ananth, R. P.; Sahoo, A. K.; Kumaravel, K.; Mohanraj, K.; Samal, A. MeFSAT: A Curated Natural Product Database Specific to Secondary Metabolites of Medicinal Fungi. *RSC Adv.* **2021**, *11*, 2596−2607.

(52) Clemons, P. A.; Bodycombe, N. E.; Carrinski, H. A.; Wilson, J. A.; Shamji, A. F.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Small Molecules of Different Origins Have Distinct Distributions of Structural Complexity That Correlate with Protein-Binding Profiles. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107*, 18787−18792.

(53) Clemons, P. A.; Wilson, J. A.; Dančík, V.; Muller, S.; Carrinski, H. A.; Wagner, B. K.; Koehler, A. N.; Schreiber, S. L. Quantifying Structure and Performance Diversity for Sets of Small Molecules Comprising Small-Molecule Screening Collections. *Proc. Natl. Acad. Sci. U. S. A.* **2011**, *108*, 6817−6822.

(54) Brown, N.; Jacoby, E. On Scaffolds and Hopping in Medicinal Chemistry. *Mini-Rev. Med. Chem.* **2006**, *6*, 1217−1229.

(55) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.

(56) Zeng, X.; Zhang, P.; Wang, Y.; Qin, C.; Chen, S.; He, W.; Tao, L.; Tan, Y.; Gao, D.; Wang, B.; Chen, Z.; Chen, W.; Jiang, Y. Y.; Chen, Y. Z. CMAUP: A Database of Collective Molecular Activities of Useful Plants. *Nucleic Acids Res.* **2019**, *47*, D1118−D1127.

(57) Sorokina, M.; Merseburger, P.; Rajan, K.; Yirik, M. A.; Steinbeck, C. COCONUT Online: Collection of Open Natural Products Database. *J. Chem.* **2021**, *13*, 2.

(58) Ntie-Kang, F.; Telukunta, K. K.; Döring, K.; Simoben, C. V.; Moumbock, A. A. F.; Malange, Y. I.; Njume, L. E.; Yong, J. N.; Sippl, W.; Günther, S. NANPDB: A Resource for Natural Products from Northern African Sources. *J. Nat. Prod.* **2017**, *80*, 2067−2076.

(59) van Santen, J. A.; Poynton, E. F.; Iskakova, D.; McMann, E.; Alsup, T. A.; Clark, T. N.; Fergusson, C. H.; Fewer, D. P.; Hughes, A. H.; McCadden, C. A.; Parra, J.; Soldatou, S.; Rudolf, J. D.; Janssen, E. M.-L.; Duncan, K. R.; Linington, R. G. The Natural Products Atlas 2.0: A Database of Microbially-Derived Natural Products. *Nucleic Acids Res.* **2022**, *50*, D1317−D1323.

(60) Banerjee, P.; Erehman, J.; Gohlke, B.-O.; Wilhelm, T.; Preissner, R.; Dunkel, M. Super Natural II−a Database of Natural Products. *Nucleic Acids Res.* **2015**, *43*, D935−D939.

(61) Wishart, D. S.; Feunang, Y. D.; Guo, A. C.; Lo, E. J.; Marcu, A.; Grant, J. R.; Sajed, T.; Johnson, D.; Li, C.; Sayeeda, Z.; Assempour, N.; Iynkkaran, I.; Liu, Y.; Maciejewski, A.; Gale, N.; Wilson, A.; Chin, L.; Cummings, R.; Le, D.; Pon, A.; Knox, C.; Wilson, M. DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* **2018**, *46*, D1074−D1082.

(62) Medina-Franco, J. L.; Martínez-Mayorga, K.; Bender, A.; Scior, T. Scaffold Diversity Analysis of Compound Data Sets Using an Entropy-Based Measure. *QSAR Comb. Sci.* **2009**, *28*, 1551−1560.

(63) González-Medina, M.; Owen, J. R.; El-Elimat, T.; Pearce, C. J.; Oberlies, N. H.; Figueroa, M.; Medina-Franco, J. L. Scaffold Diversity of Fungal Metabolites. *Front. Pharmacol.* **2017**, *8*, 180.

(64) Krier, M.; Bret, G.; Rognan, D. Assessing the Scaffold Diversity of Screening Libraries. *J. Chem. Inf. Model.* **2006**, *46*, 512−524.

(65) Ertl, P.; Rohde, B. The Molecule Cloud - Compact Visualization of Large Collections of Molecules. *J. Chem.* **2012**, *4*, 12.

(66) *Scopy*; 2022. https://scopy.iamkotori.com/ (accessed 2022-03-31).

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

(68) 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 Settings. *Adv. Drug Delivery Rev.* **2001**, *46*, 3−26.

(69) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55−68.

(70) Veber, D. F.; Johnson, S. R.; Cheng, H.-Y.; Smith, B. R.; Ward, K. W.; Kopple, K. D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615−2623.

(71) Egan, W. J.; Merz, K. M., Jr.; Baldwin, J. J. Prediction of Drug Absorption Using Multivariate Statistics. *J. Med. Chem.* **2000**, *43*, 3867−3877.

(72) Hughes, J. D.; Blagg, J.; Price, D. A.; Bailey, S.; DeCrescenzo, G. A.; Devraj, R. V.; Ellsworth, E.; Fobian, Y. M.; Gibbs, M. E.; Gilles, R. W.; Greene, N.; Huang, E.; Krieger-Burke, T.; Loesel, J.; Wager, T.; Whiteley, L.; Zhang, Y. Physiochemical Drug Properties Associated with in Vivo Toxicological Outcomes. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 4872−4875.

(73) Gleeson, M. P. Generation of a Set of Simple, Interpretable ADMET Rules of Thumb. *J. Med. Chem.* **2008**, *51*, 817−834.

(74) Conway, J. R.; Lex, A.; Gehlenborg, N. UpSetR: An R Package for the Visualization of Intersecting Sets and Their Properties. *Bioinformatics* **2017**, *33*, 2938−2940.

(75) Bickerton, G. R.; Paolini, G. V.; Besnard, J.; Muresan, S.; Hopkins, A. L. Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* **2012**, *4*, 90−98.

(76) Shannon, P.; Markiel, A.; Ozier, O.; Baliga, N. S.; Wang, J. T.; Ramage, D.; Amin, N.; Schwikowski, B.; Ideker, T. Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Res.* **2003**, *13*, 2498−2504.

(77) Schmidt, R.; Ehmki, E. S. R.; Ohm, F.; Ehrlich, H.-C.; Mashychev, A.; Rarey, M. Comparing Molecular Patterns Using the Example of SMARTS: Theory and Algorithms. *J. Chem. Inf. Model.* **2019**, *59*, 2560−2571.

(78) Ehmki, E. S. R.; Schmidt, R.; Ohm, F.; Rarey, M. Comparing Molecular Patterns Using the Example of SMARTS: Applications and Filter Collection Analysis. *J. Chem. Inf. Model.* **2019**, *59*, 2572−2586.

(79) *Flowers of India*; 2022. http://www.flowersofindia.net/ (accessed 2022-03-31).

(80) *Tropicos.org*; 2022. https://www.tropicos.org/ (accessed 2022-03-31).

(81) *Encyclopedia of Indian medicinal plants from FRLHT*; FRLHT 2022. http://envis.frlht.org/ (accessed 2022-03-31).

(82) *Medicinal Plants Names Service*; 2022. https://mpns.science.kew.org/ (accessed 2022-03-31).

(83) *International Plant Names Index*; 2022. https://www.ipni.org/ (accessed 2022-03-31).

(84) *Plants of the World Online*; 2022. https://powo.science.kew.org/ (accessed 2022-03-31).

(85) *World Flora Online*; 2022. http://www.worldfloraonline.org/ (accessed 2022-03-31).

(86) *GardenersWorld.com.*; 2022. https://www.gardenersworld.com/ (accessed 2022-03-31).

(87) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Chem.* **2011**, *3*, 33.

(88) *RDKit: Open-source cheminformatics*; 2021. https://www.rdkit.org/ (accessed 2021-01-12).

(89) Chambers, J.; Davies, M.; Gaulton, A.; Hersey, A.; Velankar, S.; Petryszak, R.; Hastings, J.; Bellis, L.; McGlinchey, S.; Overington, J. P. UniChem: A Unified Chemical Structure Cross-Referencing and Identifier Tracking System. *J. Chem.* **2013**, *5*, 3.

(90) Daina, A.; Michielin, O.; Zoete, V. SwissADME: A Free Web Tool to Evaluate Pharmacokinetics, Drug-Likeness and Medicinal Chemistry Friendliness of Small Molecules. *Sci. Rep.* **2017**, *7*, 42717.

(91) Yap, C. W. PaDEL-Descriptor: An Open Source Software to Calculate Molecular Descriptors and Fingerprints. *J. Comput. Chem.* **2011**, *32*, 1466−1474.

(92) Tweedie, S.; Braschi, B.; Gray, K.; Jones, T. E. M.; Seal, R. L.; Yates, B.; Bruford, E. A. Genenames.Org: The HGNC and VGNC Resources in 2021. *Nucleic Acids Res.* **2021**, *49*, D939−D946.

(93) Du, H.; Gao, J.; Weng, G.; Ding, J.; Chai, X.; Pang, J.; Kang, Y.; Li, D.; Cao, D.; Hou, T. CovalentInDB: A Comprehensive Database Facilitating the Discovery of Covalent Inhibitors. *Nucleic Acids Res.* **2021**, *49*, D1122−D1129.

(94) Gao, M.; Moumbock, A. F. A.; Qaseem, A.; Xu, Q.; Günther, S. CovPDB: A High-Resolution Coverage of the Covalent Protein-Ligand Interactome. *Nucleic Acids Res.* **2022**, *50*, D445−D450.

(95) Rogers, D.; Hahn, M. Extended-Connectivity Fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742−754.

(96) Franz, M.; Lopes, C. T.; Huck, G.; Dong, Y.; Sumer, O.; Bader, G. D. Cytoscape.Js: A Graph Theory Library for Visualisation and Analysis. *Bioinformatics* **2016**, *32*, 309−311.

(97) Bienfait, B.; Ertl, P. JSME: A Free Molecule Editor in JavaScript. *J. Chem.* **2013**, *5*, 24.

(98) *Jmol: an open-source Java viewer for chemical structures in 3D*; 2022. https://jmol.sourceforge.net/ (accessed 2022-03-31).