


 Cite this: *RSC Adv.*, 2021, **11**, 2596

MeFSAT: a curated natural product database specific to secondary metabolites of medicinal fungi[†]

 R. P. Vivek-Ananth, ^{‡,ab} Ajaya Kumar Sahoo, ^{‡,ab} Kavyaa Kumaravel, ^{‡,a} Karthikeyan Mohanraj, ^{ac} and Areejit Samal, ^{*ab}

Fungi are a rich source of secondary metabolites which constitutes a valuable and diverse chemical space of natural products. Medicinal fungi have been used in traditional medicine to treat human ailments for centuries. To date, there is no devoted resource on secondary metabolites and therapeutic uses of medicinal fungi. Such a dedicated resource compiling dispersed information on medicinal fungi across published literature will facilitate ongoing efforts towards natural product based drug discovery. Here, we present the first comprehensive manually curated database on Medicinal Fungi Secondary metabolites And Therapeutics (MeFSAT) that compiles information on 184 medicinal fungi, 1830 secondary metabolites and 149 therapeutics uses. Importantly, MeFSAT contains a non-redundant *in silico* natural product library of 1830 secondary metabolites along with information on their chemical structures, computed physicochemical properties, drug-likeness properties, predicted ADMET properties, molecular descriptors and predicted human target proteins. By comparing the physicochemical properties of secondary metabolites in MeFSAT with other small molecules collections, we find that fungal secondary metabolites have high stereochemical complexity and shape complexity similar to other natural product libraries. Based on multiple scoring schemes, we have filtered a subset of 228 drug-like secondary metabolites in MeFSAT database. By constructing and analyzing chemical similarity networks, we show that the chemical space of secondary metabolites in MeFSAT is highly diverse. The compiled information in MeFSAT database is openly accessible at: <https://cb.imsc.res.in/mefsat/>.

 Received 7th December 2020
 Accepted 4th January 2021

DOI: 10.1039/d0ra10322e

rsc.li/rsc-advances

Introduction

Fungi are present in every ecological niche, and thus, face challenges from myriad biotic and abiotic stressors.^{1,2} Investigation of the fungal habitats has shed important insights into the stimulation and the biosynthesis of valuable secondary metabolites produced by fungi.³ As a rich source of secondary metabolites, fungi are valuable contributors to the chemical diversity of the natural product space. Importantly, the fungal secondary metabolome is enriched in bioactive molecules and has immense potential for drug discovery⁴ including antibiotics. Notably, the first broad-spectrum antibiotic, penicillin, discovered by Alexander Fleming is a fungal secondary metabolite.

Secondary metabolites are small organic compounds mainly produced by plants, fungi and bacteria, and these metabolites are not essential for the growth and reproduction of the organism.⁵ As a rich source of bioactive molecules, it is not surprising that natural products have played an indomitable role in the history of drug discovery.⁶ By one estimate,⁶ natural products have contributed to the discovery of ~35% drugs approved by the US Food and Drug Administration (FDA) to date. Therefore, significant research effort has been devoted towards development of natural product databases^{7–15} to facilitate computational approaches to drug discovery.

In this context, several natural product databases^{8,9,11–14,16} dedicated to secondary metabolites or phytochemicals of medicinal plants have been built to date. For instance, some of us have built the IMPPAT¹⁴ database which is the largest dedicated resource on phytochemicals of Indian medicinal plants to date. During manual curation of the IMPPAT¹⁴ database, we realized that there is no dedicated resource on secondary metabolites of medicinal fungi (or mushrooms) to date. This is surprising given medicinal mushrooms^{17–21} have also been used for centuries in traditional medicine, especially traditional Chinese medicine, to treat human ailments. Therefore, a comprehensive and dedicated resource on secondary

^aThe Institute of Mathematical Sciences (IMSc), Chennai 600113, India. E-mail: asamal@imsc.res.in

^bHomi Bhabha National Institute (HBNI), Mumbai 400094, India

^cInstitute for Clinical Chemistry and Laboratory Medicine, Technische Universität Dresden, Dresden 01307, Germany

[†] Electronic supplementary information (ESI) available: Tables S1–S10 and Fig. S1 and S2. See DOI: [10.1039/d0ra10322e](https://doi.org/10.1039/d0ra10322e)

[‡] RPV, AKS and KK contributed equally to this work and should be considered as joint-first authors.



metabolites of medicinal fungi along with their therapeutic uses will serve ongoing research efforts towards natural product based drug discovery. Here, we have addressed this unmet need by building a natural product database dedicated to secondary metabolites of medicinal fungi.

The fungal kingdom is very large and encompasses diverse organisms ranging from simple yeasts to mushrooms. Some fungi are considered 'medicinal' due to the beneficial bioactivity of their secondary metabolites and/or their usage in systems of traditional medicine to treat human ailments.¹⁷⁻²¹ Mushrooms are macrofungi with fruiting bodies²² that have been used as food and/or medicine for centuries across many civilizations.^{17-21,23} Mushroom-derived preparations have been used in traditional medicine practiced in many Asian countries.²⁴ Presently, the valuable information on secondary

metabolites and therapeutic uses of medicinal fungi is dispersed across published literature including articles and books,¹⁷⁻²¹ and this limits its effective use for drug discovery. Moreover, existing microbial natural product databases such as NPATLAS¹⁵ have certain limitations; they are neither specific to fungi nor do they capture both secondary metabolite and therapeutic use information for medicinal fungi. In other words, a common repository of high-quality information on secondary metabolites and therapeutic uses of medicinal fungi is needed to harness the potential of this chemical space for drug discovery.

In this direction, we present a manually curated database, Medicinal Fungi Secondary metabolites And Therapeutics (MeFSAT), dedicated to secondary metabolites and therapeutic

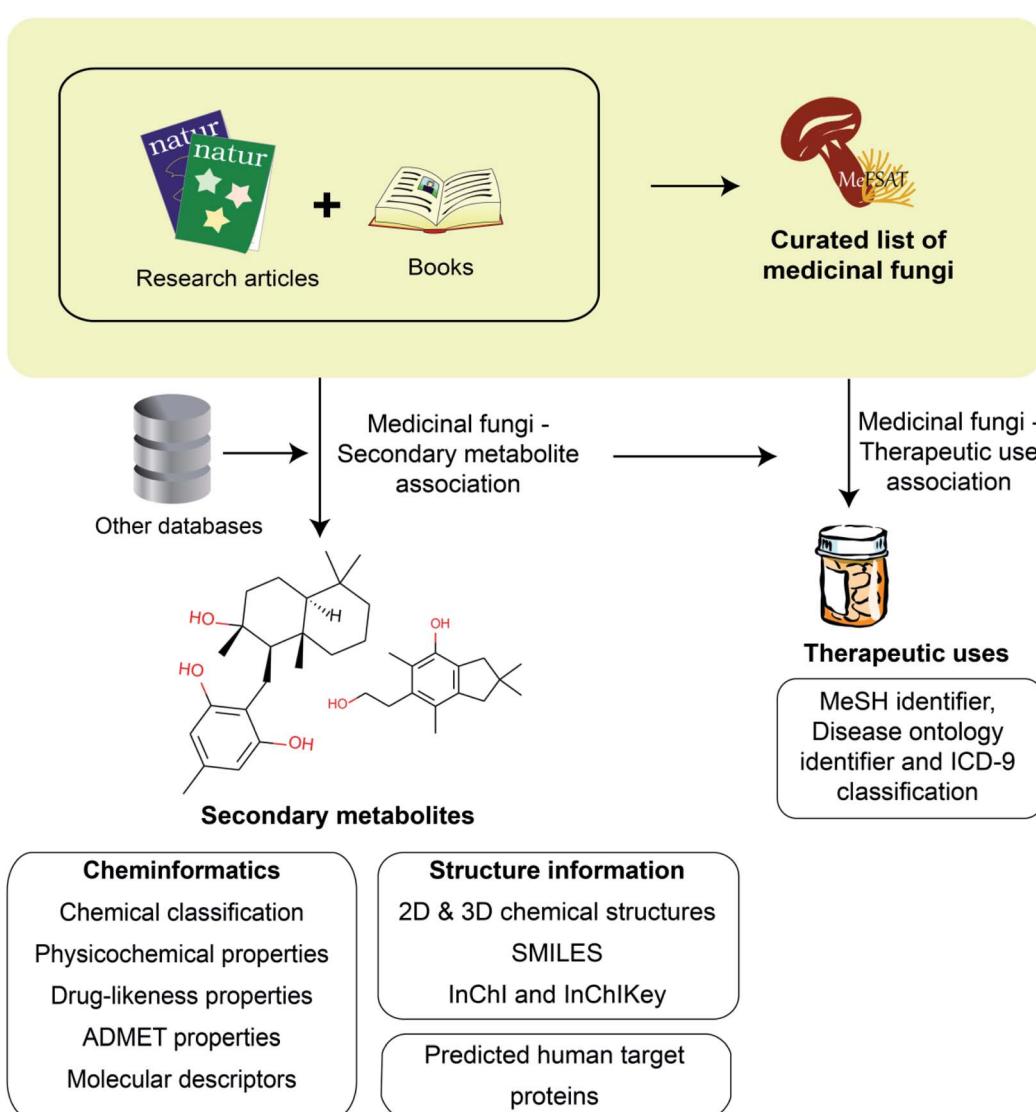


Fig. 1 Schematic overview of the workflow to construct the MeFSAT database. Briefly, we compiled a curated list of medicinal fungi from the published literature. Next, we mined the published literature to compile secondary metabolites produced by different medicinal fungi. For the manually curated non-redundant list of secondary metabolites produced by medicinal fungi, we have compiled the chemical structures and employed cheminformatics tools to compute their physicochemical, drug-likeness and ADMET properties. Subsequently, we have compiled and curated information on the therapeutic uses of medicinal fungi from the published literature.



uses of medicinal fungi which is openly accessible at: <https://cb.imsc.res.in/mefsat/>.

Methods

Workflow for the compilation and curation of secondary metabolites of medicinal fungi

Fungal secondary metabolites are a valuable chemical space of diverse natural products with several applications including in drug discovery.²⁵ Medicinal Fungi Secondary metabolites And Therapeutics (MeFSAT) is a manually curated database that compiles information on secondary metabolites and reported therapeutic uses of medicinal fungi from published research articles and specialized books on the subject. In the following subsections, we provide an overview of the steps involved in the construction of MeFSAT database (Fig. 1).

Curated list of medicinal fungi. The first step in the database construction workflow involved the compilation of a curated list of medicinal fungi from published literature. For this purpose, we performed an extensive PubMed²⁶ search using the query “Medicinal fungi” OR “Medicinal mushroom”, and this keyword search last performed on 3 May 2020 led to the retrieval of 1206 published research articles (ESI Table S1†). Apart from research articles, we also curated information on medicinal fungi from books^{17–21} on the topic. In the first-pass, we obtained a list of 354 fungi names with medicinal use from published literature consisting of research articles and books. Since the use of synonymous fungal names is common in the published literature, we next mapped the 354 fungal names with medicinal use to their accepted names using two resources namely, Catalogue of Life: 2019 Annual Checklist²⁷ and Mycobank.²⁸ In the end, this mapping to accepted names led to a non-redundant list of 253 medicinal fungi (ESI Table S2†). Note that the curated list of 253 medicinal fungi does not include Lichens.²⁹

Secondary metabolites of medicinal fungi. The second step in our database construction workflow involved the search for secondary metabolites of 253 medicinal fungi compiled from the published literature (ESI Table S2†). From research articles and books,^{17–21,30} we were able to gather information on 1139 secondary metabolites with evidence of being produced by at least one of 145 medicinal fungi. Further, we were able to gather additional information on 892 and 7 secondary metabolites with evidence of being produced by at least one of 121 and 5 medicinal fungi from two microbial natural product databases, namely, NPATLAS¹⁵ and novel Antibiotics database (<http://www.antibiotics.or.jp/journal/database/database-top.htm>), respectively. Overall, we were able to gather chemical names of 2038 secondary metabolites with evidence of being produced by at least one of 188 medicinal fungi from above-mentioned sources.

Non-redundant *in silico* library of secondary metabolites of medicinal fungi. The primary objective of MeFSAT is to build a non-redundant curated resource of secondary metabolites of medicinal fungi along with information on their two-dimensional (2D) and three-dimensional (3D) chemical structure. Towards this objective, the compiled information on

chemical names of secondary metabolites of medicinal fungi was evaluated to create a structurally non-redundant *in silico* chemical library. Specifically, the chemical names of the compiled secondary metabolites were mapped to chemical identifiers employed by standard chemical databases, namely, PubChem,³¹ ChemIDplus (<https://chem.nlm.nih.gov/chemidplus/>), ChemSpider³² and NPATLAS.¹⁵ In case, we were unable to map a secondary metabolite to an identifier in at least one of the above-mentioned chemical databases, the chemical structure of the secondary metabolite was manually drawn from the corresponding published research article. Following the above-mentioned steps, we were able to obtain the chemical structure information for 1991 secondary metabolites which have evidence of being produced by at least one of 184 medicinal fungi. Note that a few secondary metabolites were omitted from further consideration as we were unable to either map them to an identifier or obtain their chemical structure from published literature.

Thereafter, we used an in-house Python script which employs Tanimoto coefficient³³ (T_c) to determine chemical similarity between secondary metabolites. To create a non-redundant chemical library, we have merged compiled secondary metabolites from published literature if they were determined to be identical based on their chemical structure. Importantly, our manual curation effort while creating a non-redundant database of secondary metabolites has also taken into due consideration the stereochemistry of the chemical structures. Finally, this effort has led to a non-redundant set of 1830 secondary metabolites in MeFSAT database with literature evidence of being produced by at least one of 184 medicinal fungi (ESI Tables S3 and S4†).

We remark that our primary objective is to create a resource on secondary metabolites of medicinal fungi rather than secondary metabolites of any fungi. Therefore, we decided to first mine literature to compile a curated list of medicinal fungi from published literature, and thereafter, performed literature search to compile known secondary metabolites produced by any of the medicinal fungi in the curated list.

Annotation of secondary metabolites of medicinal fungi. For the 1830 secondary metabolites in MeFSAT database, the 2D chemical structures were saved in SDF, MOL and MOL2 file formats using OpenBabel³⁴ while their 3D chemical structures were manually retrieved from PubChem³¹ if available. For the remaining secondary metabolites whose 3D structures are not available in PubChem, the 3D structures were generated using RDKit³⁴ (<http://www.rdkit.org/>) by embedding the molecule using ETKDG method³⁵ followed by energy minimization using MMFF94 force field.³⁶ The 3D structures for secondary metabolites were saved in SDF, MOL, MOL2, PDB and PDBQT file formats using OpenBabel.³⁷ Apart from the 2D and 3D structure information, the SMILES, InChI and InChIKey of the secondary metabolites were also generated using OpenBabel³⁷ (ESI Table S5†). Moreover, the secondary metabolites in MeFSAT were hierarchically classified into chemical Kingdom, chemical SuperClass, chemical Class and chemical SubClass using ClassyFire³⁸ (<http://classyfire.wishartlab.com/>) (ESI Table S5†).



The basic physicochemical properties of the secondary metabolites in MeFSAT database were computed using RDKit³⁴ and SwissADME³⁹ (<http://www.swissadme.ch/>) (ESI Table S6†). To assess the drug-likeness of the secondary metabolites in MeFSAT, we have computed multiple scoring schemes and properties namely, Lipinski's rule of five (RO5),⁴⁰ Ghose filter,⁴¹ Veber filter,⁴² Egan filter,⁴³ Pfizer's 3/75 filter,⁴⁴ GlaxoSmithKline's (GSK) 4/400,⁴⁵ number of leadlikeness violations⁴⁶ and weighted quantitative estimate of drug-likeness (QEDw)⁴⁷ using RDKit³⁴ and SwissADME³⁹ (ESI Table S7†). The assessment of Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) properties is essential in the drug discovery pipeline. We have used SwissADME³⁹ to predict the ADMET properties of the secondary metabolites in MeFSAT (ESI Table S8†). Finally, we have computed 1875 (2D and 3D) molecular descriptors for each secondary metabolite in MeFSAT using PaDEL⁴⁸ (<http://padel.nus.edu.sg/software/padeldescriptor>). These molecular descriptors can be broadly categorized into different classes such as chemical composition, topology, 3D shape and functionality.

Genome sequences of medicinal fungi. For the 184 medicinal fungi that have secondary metabolite information in MeFSAT, we have gathered their genome sequencing status from the Joint Genome Institute (JGI) portal⁴⁹ (<https://genome.jgi.doe.gov/>) and the National Center for Biotechnology Information (NCBI)²⁶ (<https://www.ncbi.nlm.nih.gov/genome>). In addition, we have compiled information from published literature on the traditional system of medicine in which these medicinal fungi are used.

Therapeutic uses of medicinal fungi. The next step in the database construction workflow involved the compilation of therapeutic uses of medicinal fungi from published literature including specialized books.^{20,21} Note that the compiled therapeutic uses are based on available information on the use of medicinal fungi to treat human diseases. Notably, we have manually curated the compiled therapeutic use terms from various literature sources to create a non-redundant list of standardized therapeutic use terms for medicinal fungi in MeFSAT by mapping the therapeutic use terms to standard identifiers from Medical Subject Headings (MeSH),⁵⁰ Disease Ontology⁵¹ and ICD-9-CM chapters.⁵² In sum, this effort has led to a non-redundant list of 149 standardized therapeutic use terms in MeFSAT that are associated with different medicinal fungi (ESI Table S9†). Note that MeFSAT compiles therapeutic uses at the level of medicinal fungi rather than secondary metabolites from published literature.

Predicted human target proteins of secondary metabolites. Lastly, we have compiled the predicted human target proteins of secondary metabolites in MeFSAT from the STITCH⁵³ database (<http://stitch.embl.de/>). To date, STITCH⁵³ is the largest resource on predicted interactions between chemicals and their target proteins. In MeFSAT database, we have included only high confidence interactions between secondary metabolites and human target proteins that have a combined STITCH⁵³ score ≥ 700 . Further, we have also mapped the genes corresponding to predicted human target proteins of secondary

metabolites from STITCH⁵³ to their respective HUGO Gene Nomenclature Committee (HGNC) symbols.⁵⁴

Comparison of physicochemical properties of fungal secondary metabolites with other small molecule collections

We have compared the physicochemical properties of secondary metabolites in MeFSAT with three other small molecule collections studied by Clemons *et al.*⁵⁵ The three small molecule libraries are: (a) a library of commercial compounds (CC) containing 6152 representative small molecules from commercial sources, (b) a library of diversity-oriented synthesis compounds (DC') containing 5963 small molecules synthesized by the academic community, and (c) a library of natural products (NP) containing 2477 small molecules from various natural sources including microbes and plants. Note that the set of 1830 secondary metabolites in MeFSAT is also a natural product library, however, there is only a tiny overlap of 20 small molecules between NP library of Clemons *et al.*⁵⁵ and secondary metabolites in our database. Further, the computation of physicochemical properties failed for 3 small molecules each in CC and DC' libraries of Clemons *et al.*⁵⁵ and thus, we have omitted them from later analysis.

Chemical similarity of secondary metabolites

Tanimoto coefficient³³ (T_c) is a widely used metric to compute chemical structure similarity,⁵⁶ and we have used T_c based on Extended Circular Fingerprints (ECFP4)⁵⁷ to compute the structure similarity between secondary metabolites in MeFSAT and small molecule drugs approved by US FDA. For this purpose, the structures of FDA approved drugs were retrieved from DrugBank.⁵⁸ Note that the computed T_c value between any two molecules has a range between 0 and 1, wherein 0 represents little or no structure similarity and 1 represents very high or exact structure similarity. Based on a previous study by Jaisal *et al.*,⁵⁹ we have chosen the cutoff of $T_c \geq 0.5$ to decide if a given pair of chemicals have significant structure similarity.

Chemical similarity network

We have constructed chemical similarity networks (CSNs) wherein nodes are chemicals and edges between pairs of chemicals signify high chemical similarity. For this construction, T_c based on ECFP4 (ref. 57) gives the extent of similarity between two chemicals, and thus, the weights of edges between nodes in the network. In the CSN, we only retain edges between two chemicals if the T_c between them is ≥ 0.5 .

We have constructed two CSNs here that correspond to 1830 secondary metabolites in MeFSAT and the subset of 228 drug-like secondary metabolites in MeFSAT, respectively. Further, these CSNs were visualized using Cytoscape.⁶⁰

Maximum common substructure

Maximum Common Substructure (MCS) of two or more chemical structures is the largest common substructure that is present in them. MCS has many applications including in chemical similarity search and hierarchical clustering of



(a)
Boletus edulis

Family: Boletaceae

System of Traditional medicine: Traditional Chinese Medicine

Reference: PMID:30668382, ISBN:1-57067-143-5

Genome sequence

| Medicinal fungi | Secondary metabolite identifier | Secondary metabolite name | Reference |
|-----------------|---------------------------------|------------------------------------|---------------|
| Boletus edulis | MSID0000090 | (1S)-3-Ethoxyphenyl-1,2-ethanediol | PMID:17440922 |
| Boletus edulis | MSID0000092 | (1S)-4-Ethoxyphenyl-1,2-ethanediol | PMID:17440922 |
| Boletus edulis | MSID0000124 | 1-(3-Ethoxyphenyl)-1,2-ethanediol | PMID:17440922 |
| Boletus edulis | MSID0000125 | 1-(3-Formylphenyl)-1,2-ethanone | PMID:17440922 |
| Boletus edulis | MSID0000126 | 1-(4-Ethoxyphenyl)-1,2-ethanediol | PMID:17440922 |

(c)

| Medicinal fungi | Therapeutic use | Therapeutic use identifier | Reference |
|-----------------|-----------------------|--|---------------|
| Boletus edulis | Antineoplastic agents | MESH:D000970, MESH:D000969, ICD-9-199.1, ICD-9-239.9, ICD-9-E933.1; Doid:162 | PMID:30668382 |
| Boletus edulis | Generalized pain | ICD-9-780.96 | PMID:30668382 |
| Boletus edulis | Stomach pain | PMID:30668382 | |

(d)

Physicochemical filter Drug-like filter Chemical similarity filter

Molecular Weight LogP Topological Polar Surface Area (TPSA)

Hydrogen Bond Acceptors (HBA) Hydrogen Bond Donors (HBD) Heavy Atoms

Heteroatoms Number of Rings Rotatable Bonds

Stereochemical Complexity Shape Complexity

Search

Physicochemical filter Drug-like filter Chemical similarity filter

Lipinski ROS Violation Ghose Rule Violation GSK 4/400

Ptzer 3/75 Veber Rule Egan Rule

QEDw

Search

Physicochemical filter Drug-like filter Chemical similarity filter

Enter SMILES Choose Fingerprint

Search

(b)

Secondary metabolite: Trametenolic acid B

Summary Physicochemical properties Drug-likeness properties ADMET properties Descriptors Predicted human target proteins

Chemical structure

Chemical structure download

2D: 2D-MOL 2D-MOL2 2D-PDB 2D-PDBQT

3D: 3D-MOL 3D-MOL2 3D-PDB 3D-PDBQT

Physicochemical properties

| Property name | Tool | Property value |
|--|-------|----------------|
| Molecular weight (g/mol) | RDKit | 456.71 |
| Log P | RDKit | 7.54 |
| Topological polar surface area (Å ²) | RDKit | 57.53 |
| Number of hydrogen bond acceptors | RDKit | 2 |
| Number of hydrogen bond donors | RDKit | 2 |

Drug-likeness properties

| Property name | Tool | Property value |
|--|-----------|----------------|
| Number of Lipinski's rule of 5 violations | RDKit | 1 |
| Lipinski's rule of 5 filter | RDKit | Passed |
| Number of Ghose filter violations | RDKit | 3 |
| Ghose filter | RDKit | Failed |
| Veber filter | RDKit | Good |
| Egan filter | RDKit | Bad |
| Ptzer's 3/75 filter | RDKit | Bad |
| GSK 4/400 filter | RDKit | Bad |
| Number of Leadlikeness violations | SwissADME | 2 |
| Weighted quantitative estimate of drug-likeness (QEDw) score | RDKit | 0.42 |

ADMET properties

| Property name | Tool | Property value |
|--|-----------|----------------|
| Bioavailability score | SwissADME | 0.85 |
| Solubility class [ESOL] | SwissADME | Poorly soluble |
| Solubility class [Silicos-IT] | SwissADME | Poorly soluble |
| Blood Brain Barrier permeation | SwissADME | No |
| Gastrointestinal absorption | SwissADME | Low |
| Log K _{o/w} (Skin permeation, cm ² /s) | SwissADME | -3.78 |
| Number of PAHs structural alerts | SwissADME | 0 |
| Number of Breyer structural alerts | SwissADME | 1 |

Descriptors

| Tool | Type | Descriptor | Description | Descriptor class | Result |
|-------|------|------------|--|------------------|----------|
| PaDEL | 2D | nAdd | Number of acidic groups. The list of acidic groups is defined by these SMARTS "[S](=O)(=O)[C]([S]([O-])=O]", "[S]([O-])=O", "[S]([O-])=O", "[S]([O-])=O", and "[S]([O-])=O". | Acidic Group | 1 |
| PaDEL | 2D | ALogP | Ghose-Crippen LogK _{o/w} | ALogP descriptor | 3.3496 |
| PaDEL | 2D | ALogP2 | Square of ALogP | ALogP descriptor | 11.2198 |
| PaDEL | 2D | AMR | Molar refractivity | ALogP descriptor | 135.1754 |

Predicted human target proteins

Trametenolic acid B

TYR 800

Fig. 2 Web interface of the MeFSAT database. (a) Snapshot of the result of a standard query for secondary metabolites of a medicinal fungus. The example shows the secondary metabolites for the fungus *Boletus edulis*. (b) Snapshot of the detailed information page for a secondary metabolite which gives its 2D and 3D chemical structure, physicochemical properties, drug-likeness properties, predicted ADMET properties, molecular descriptors and predicted human target proteins. From this page, users can download the 2D and 3D structure of the secondary metabolite in SDF, MOL, MOL2, PDB or PDBQT file formats. The example shows information for the secondary metabolite trametenolic acid B. (c) Snapshot of the result of a standard query for therapeutic uses of a medicinal fungus. The example shows the therapeutic uses for the fungus *Boletus edulis*. (d) Snapshot of the advanced search options which enable users to filter secondary metabolites based on their physicochemical properties or drug-likeness properties or chemical similarity with a query chemical structure in SMILES format.

2600 | RSC Adv., 2021, 11, 2596–2607

© 2021 The Author(s). Published by the Royal Society of Chemistry

Open Access Article. Published on 12 January 2021. Downloaded on 1/18/2026 10:13:46 PM.
 This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

chemicals.⁶¹ We have used RDKit³⁴ to identify the MCSs for sets of secondary metabolites that form different clusters in the CSN comprising of 228 drug-like secondary metabolites in MeFSAT. Note that the MCSs were only computed for the 10 largest connected components (or clusters) in the CSN of drug-like secondary metabolites. Further, the identified MCSs were visualized using the online web-server SMARTSview⁶² (<https://smartsview.zbh.uni-hamburg.de/>).

Web interface and database management system

MeFSAT database compiles information mainly on two types of association, (a) medicinal fungi and their secondary metabolites (ESI Table S4†), and (b) medicinal fungi and their therapeutic uses (ESI Table S9†). In addition, MeFSAT compiles detailed information for secondary metabolites such as their 2D and 3D chemical structure, chemical classification, physicochemical properties, drug-likeness properties, predicted ADMET properties, molecular descriptors and predicted human target proteins (ESI Tables S5–S8†). MeFSAT is openly accessible at: <https://cb.imsc.res.in/mefsat/>.

The compiled information in MeFSAT is stored in a SQL database created using the open-source relational database management system MariaDB (<https://mariadb.org/>). The web interface of MeFSAT was created using the free and open-source CSS framework Bootstrap 4.1.3 (<https://getbootstrap.com/docs/4.0/getting-started/introduction/>), with custom HTML, PHP (<http://php.net/>), CSS, JavaScript and jQuery (<https://jquery.com/>) scripts. To render visualizations in the web interface, we have used Cytoscape.js (<http://js.cytoscape.org/>) and Google Charts (<https://developers.google.com/chart>). The MeFSAT website is hosted on an Apache server (<https://httpd.apache.org/>) running on Debian 9.13 Linux operating system.

Results and discussion

MeFSAT database and its web interface

In this work, we have built a natural product database, namely, Medicinal Fungi Secondary metabolites And Therapeutics (MeFSAT), which contains manually curated information on secondary metabolites and therapeutic uses of medicinal fungi from published literature (Methods; Fig. 1). The database has a modern and intuitive web interface that enables easy access to the curated information (Fig. 2). MeFSAT is openly accessible at: <https://cb.imsc.res.in/mefsat/>.

The MeFSAT web interface enables users to retrieve manually curated associations between medicinal fungi and secondary metabolites or therapeutic uses by querying for either (a) scientific names of medicinal fungi, (b) secondary metabolite identifier, (c) secondary metabolite name, or (d) therapeutic use terms (Fig. 2). The query result is displayed as a table of associations with relevant literature references. In resultant table obtained after the search for secondary metabolite associations of medicinal fungi, the users can click on a specific medicinal fungi name that will redirect them to a dedicated page containing all secondary metabolite associations for the specific

fungi (Fig. 2a). The users can also view the detailed information page for each secondary metabolite by clicking the secondary metabolite identifiers in the above-mentioned table. The detailed information page for a secondary metabolite provides a summary of the chemical structure, external database identifiers, synonymous names, chemical classification, 2D and 3D chemical structure in different file formats, physicochemical properties, drug-likeness properties, predicted ADMET properties, molecular descriptors and predicted human target proteins (Fig. 2b).

In resultant table obtained after the search for therapeutic use associations of medicinal fungi, the users can click on a specific medicinal fungi name that will redirect them to a dedicated page containing therapeutic uses of the specific fungi which were curated from published literature along with the identifiers for therapeutic use terms from MeSH,⁵⁰ disease ontology⁵¹ and ICD-9-CM chapters⁵² (Fig. 2c). Further, by clicking the therapeutic use terms in the above-mentioned table, the users can view the list of medicinal fungi that have a specific therapeutic use.

Advanced search options in MeFSAT web interface enable users to filter the secondary metabolites by either: (a) physicochemical properties, (b) drug-likeness properties, or (c) chemical structure similarity (Fig. 2d). Specifically, the physicochemical filter tab enables users to retrieve secondary metabolites with desired physicochemical properties. The drug-likeness filter tab enables users to select secondary metabolites that pass or fail multiple drug-likeness scoring schemes. Lastly, the chemical similarity filter enables users to search for 10 secondary metabolites within MeFSAT that have the highest structural similarity to a query chemical compound entered in SMILES format. The results from these advanced search options are rendered as tables wherein secondary metabolites can be sorted based on the chosen properties.

Curated information on medicinal fungi, their secondary metabolites and therapeutic uses

The MeFSAT database compiles manually curated information on 1830 secondary metabolites produced by at least one of 184 medicinal fungi (ESI Tables S3–S8†). For this curated list of 184 medicinal fungi in MeFSAT, we have compiled information on their taxonomic family, genome sequencing status, and usage in different systems of traditional medicine. Interestingly, 54 out of the 184 medicinal fungi in MeFSAT are used in traditional Chinese medicine to treat human ailments. Further, the 184 medicinal fungi are distributed across 48 taxonomic families of which the 5 families Polyporaceae, Ganodermataceae, Agaricaceae, Hymenochaetaceae and Pleurotaceae have 24, 20, 18, 13 and 10 medicinal fungi, respectively, in MeFSAT database (ESI Table S3†).

There are 2127 medicinal fungi – secondary metabolite associations in MeFSAT database which encompass 184 medicinal fungi and 1830 secondary metabolites (Fig. 1; ESI Table S4†). These 1830 secondary metabolites are distributed across 13 chemical SuperClasses as computed using ClassyFire³⁸ (Fig. 3a and ESI Table S5†). Notably, more than 60% of the



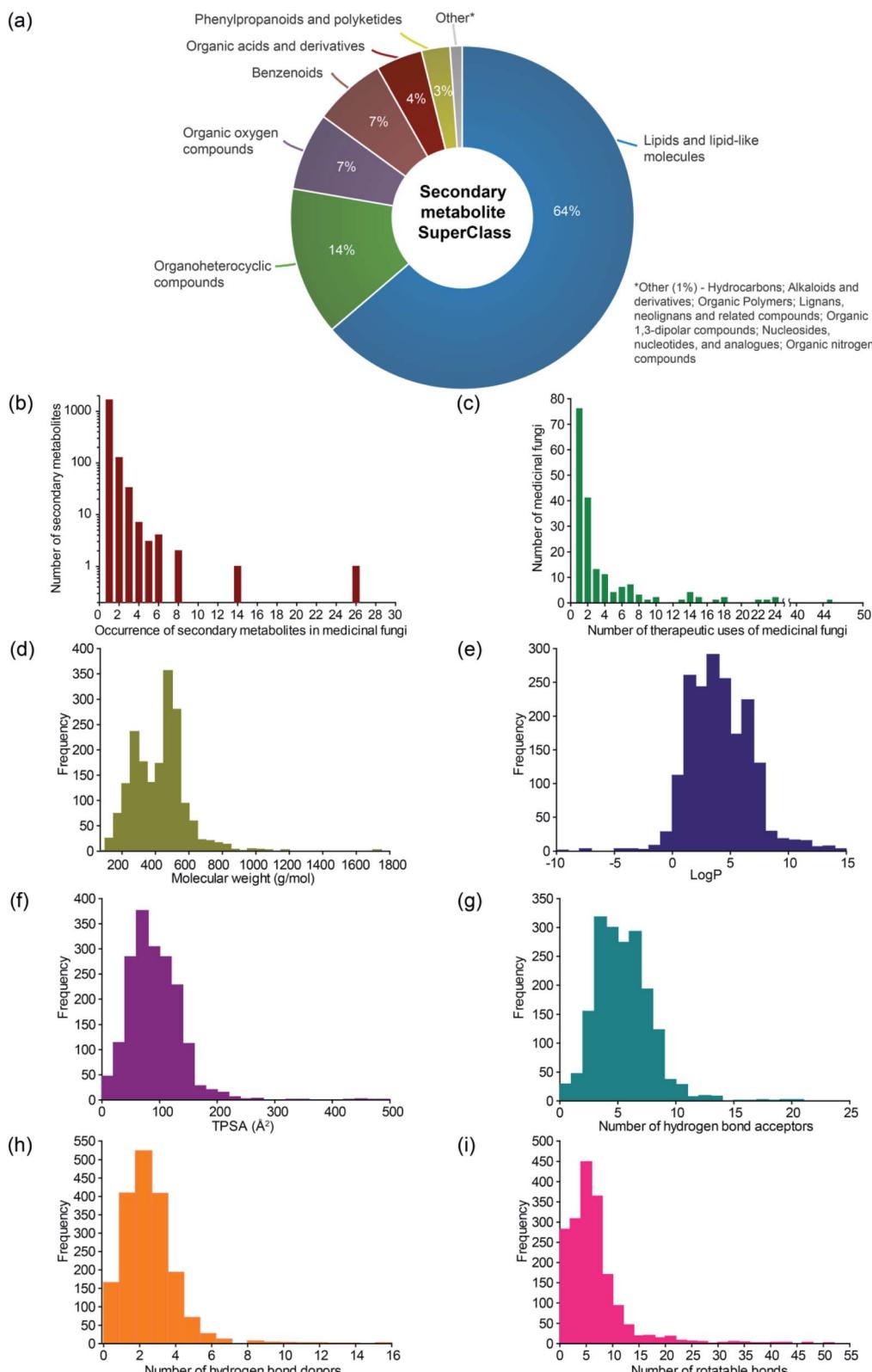


Fig. 3 Basic statistics for medicinal fungi, their secondary metabolites and therapeutic uses in MeFSAT database. (a) Pie chart shows the distribution of the secondary metabolites in MeFSAT across chemical SuperClasses obtained from ClassyFire.²¹ (b) Histogram of the number of medicinal fungi with literature evidence of producing a given secondary metabolite in MeFSAT database. (c) Histogram of the number of therapeutic uses per medicinal fungi in MeFSAT database. (d–i) Histogram of the distribution of molecular weight (g mol⁻¹), log P, topological polar surface area (TPSA) (Å²), number of hydrogen bond acceptors, number of hydrogen bond donors, and number of rotatable bonds, for the secondary metabolites in MeFSAT database.



secondary metabolites in MeFSAT belong to the chemical SuperClass 'Lipids and lipid-like molecules'. Other chemical SuperClasses enriched in secondary metabolites from MeFSAT include 'Organoheterocyclic compounds' (14%), 'Organic oxygen compounds' (7%) and 'Benzeneoids' (7%) (Fig. 3a). Among the 184 medicinal fungi, *Ganoderma lucidum* has the highest number (277) of secondary metabolite associations, followed by *Ganoderma applanatum* with 131 secondary metabolite associations, and *Hericium erinaceus* with 104 secondary metabolite associations.

In Fig. 3b, we show the histogram of the occurrence of secondary metabolites across 184 medicinal fungi in MeFSAT database. It is seen that the majority of the secondary metabolites (1779) in MeFSAT database have published literature evidence of being produced by less than 3 medicinal fungi. Further, only two secondary metabolites in MeFSAT database have published literature evidence of being produced by more than 10 medicinal fungi; these metabolites are L-ergothioneine (MSID001103) with evidence from 26 medicinal fungi and ergosterol peroxide (MSID000285) with evidence from 14 medicinal fungi. MeFSAT database also compiles 2036 predicted interactions between secondary metabolites and their human target proteins from the STITCH⁵³ database, and these interactions encompass 54 secondary metabolites and 1003 human proteins.

There are 689 medicinal fungi – therapeutic use associations in MeFSAT database which encompass 179 medicinal fungi and 149 therapeutic uses (Fig. 1 and ESI Table S9†). Fig. 3c shows the histogram of the number of therapeutic uses per medicinal fungi as compiled in MeFSAT database. It is seen that the majority of the medicinal fungi (162) in MeFSAT have less than 10 reported therapeutic uses, while 5 medicinal fungi have more than 20 reported therapeutic uses in published literature. *Ganoderma lucidum* has the highest number (45) of reported therapeutic uses, followed by *Hericium erinaceus* (24), *Lignosus rhinocerus* (24), *Antrodia cinnamomea* (23) and *Tropicoporus li-*teus (22).

Physicochemical properties of secondary metabolites in MeFSAT database and comparison with other small molecule collections

For the 1830 secondary metabolites in MeFSAT database, we have computed several physicochemical properties (ESI Table S6†). In Fig. 3d-i, we show the distribution of six physicochemical properties namely, molecular weight, log *P*, topological polar surface area (TPSA), number of hydrogen bond acceptors, number of hydrogen bond donors, and number of rotatable bonds across the 1830 secondary metabolites in MeFSAT database.

Previously, Clemons *et al.*⁵⁵ have shown that two size-independent metrics namely, stereochemical complexity and shape complexity, are excellent predictors of target protein binding specificity of small molecules. Stereochemical complexity measures the ratio of the number of chiral carbon atoms to the total number of carbon atoms in a molecule, whereas shape complexity is the ratio of the number of sp^3

hybridized carbon atoms to the total number of sp^2 and sp^3 hybridized carbon atoms in a molecule.⁵⁵ Specifically, Clemons *et al.*⁵⁵ have correlated the stereochemical and shape complexity of small molecules with their target protein binding specificity across three representative small molecule collections namely, CC, DC' and NP (Methods). Small molecules in NP collection were found to have higher stereochemical and shape complexity in comparison with those in DC' or CC collection, and moreover, small molecules in NP collection were found to be more specific binders of target proteins with low fraction of promiscuous binders in comparison with those in DC' or CC collection.⁵⁵ In other words, natural products^{14,55} were found to have higher stereochemical and shape complexity while being specific binders of target proteins.

Here, we have computed and compared the stereochemical complexity and shape complexity of the 1830 secondary metabolites in MeFSAT with those of small molecules in CC, DC' and NP collections (Fig. 4a and b). Interestingly, we find

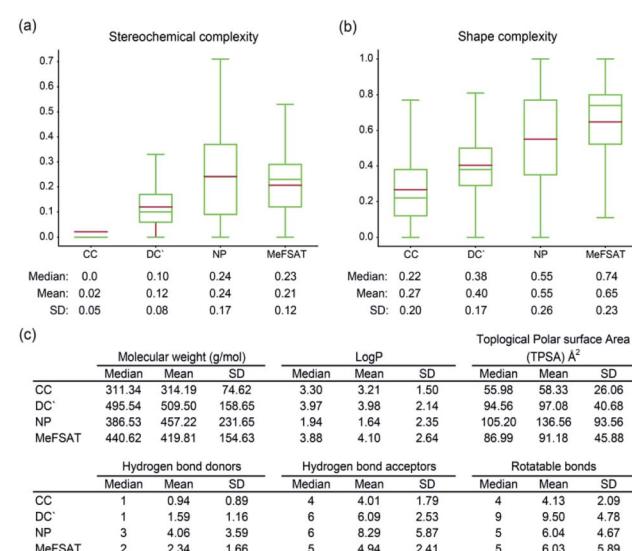


Fig. 4 Comparison of the physicochemical properties of secondary metabolites in MeFSAT with other small molecule collections. (a) Box plot shows the distribution of the stereochemical complexity of the small molecule collections CC, DC', NP and MeFSAT secondary metabolites. The median, mean and standard deviation (SD) of the stereochemical complexity for each small molecule collection is shown below the box plot. (b) Box plot shows the distribution of the shape complexity of the small molecule collections CC, DC', NP and MeFSAT secondary metabolites. The median, mean and SD of the shape complexity for each small molecule collection is shown below the box plot. Note the lower end of the box shows the first quartile, upper end of the box shows the third quartile, green line shows the median and brown line shows the mean of the distribution of stereochemical complexity or shape complexity in the two box plots. (c) Median, mean and SD of six physicochemical properties, namely, molecular weight, log *P*, topological polar surface area (TPSA), number of hydrogen bond acceptors, number of hydrogen bond donors, and number of rotatable bonds for the small molecule collections CC, DC', NP, and MeFSAT secondary metabolites. Here, CC, DC' and NP refer to small molecule collections of commercial compounds, diversity-oriented synthesis compounds and natural products, respectively (Methods).



that the mean and median of the stereochemical complexity or shape complexity of secondary metabolites in MeFSAT are closer to the NP collection than DC' or CC collections. This suggests that secondary metabolites in MeFSAT are more likely to be specific binders of target proteins than promiscuous binders. Moreover, apart from stereochemical and shape complexity, we have also compared the mean and median of six other physicochemical properties namely, molecular weight, $\log P$, TPSA, number of hydrogen bond acceptors, number of hydrogen bond donors and number of rotatable bonds, for the 1830 secondary metabolites in MeFSAT with those for small molecules in NP, DC' and CC collections (Fig. 4c).

Drug-like secondary metabolites of medicinal fungi

Natural products have directly or indirectly contributed to the discovery of $\sim 35\%$ of the small molecule drugs approved by the US FDA till 2014.⁶ To date, several scoring schemes or rules have been proposed to assess the drug-likeness of small molecules.⁶³ Here, we have used six scoring schemes namely, Lipinski's RO5,⁴⁰ Ghose filter,⁴¹ Veber filter,⁴² Egan filter,⁴³ Pfizer's 3/75 filter⁴⁴ and GSK 4/400⁴⁵ to access the drug-likeness of 1830 secondary metabolites in MeFSAT (Methods). Notably, we have identified a subset of 228 secondary metabolites ($\sim 12\%$) in MeFSAT to be drug-like, and these metabolites have passed all the six scoring schemes mentioned above. In Fig. 5a, we show the number of secondary metabolites in MeFSAT that pass different combinations of above-mentioned six scoring schemes. It is important to highlight that several natural products that failed to pass drug-likeness scores have been successfully developed into drugs.⁶⁴ Therefore, we expect a higher fraction of secondary metabolites in MeFSAT, greater than the $\sim 12\%$ metabolites that pass the six scoring schemes, have the potential to be developed into drugs. Fig. 5b shows the chemical classification of the 228 drug-like secondary metabolites that pass the six scoring schemes. The 228 drug-like secondary metabolites in MeFSAT were distributed across 8 chemical SuperClasses, with more than 30% classified as 'Organoheterocyclic compounds'. Furthermore, based on the computation of the QEDw metric, we find that 19 out of 228 drug-like secondary metabolites in MeFSAT have a high QEDw value of >0.80 (Fig. 5c). Moreover, ESI Fig. S1† shows the number of secondary metabolites within MeFSAT that pass at least 1, 2, 3, 4 or 5 out of the 6 drug-likeness scoring schemes evaluated here. It is seen that 630 secondary metabolites ($\sim 34\%$) pass at least 5 out of the 6 scoring schemes (ESI Fig. S1†). This further highlights the potential of the MeFSAT chemical space for drug discovery.

Chemical similarity networks of secondary metabolites

Chemical similarity networks (CSNs) can facilitate visualization and exploration of the structural diversity in a chemical library, and thus, enable better selection of lead compounds from a chemical space for drug development. Here, we have constructed two CSNs (Methods) wherein the first corresponds to 1830 secondary metabolites (ESI Fig. S2†) while the second corresponds to 228 drug-like secondary metabolites (Fig. 6a)

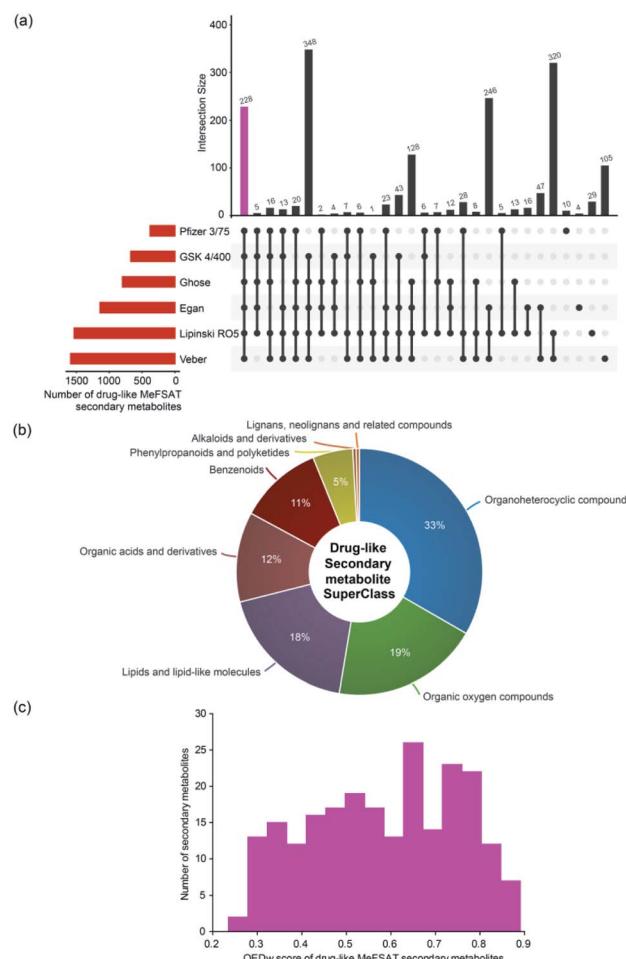


Fig. 5 Drug-likeness analysis of the secondary metabolites in MeFSAT database. (a) Evaluation of drug-likeness of secondary metabolites based on multiple scores. The horizontal bar plot shows the number of secondary metabolites in MeFSAT database that satisfy different drug-likeness scoring schemes (Methods). The vertical bar plot shows the overlap between sets of secondary metabolites that satisfy different drug-likeness scoring schemes. The pink bar in the vertical plot gives the 228 secondary metabolites that satisfy all the 6 drug-likeness scoring schemes. This plot was generated using the UpSetR package.⁶³ (b) Classification of the 228 drug-like secondary metabolites into chemical SuperClasses obtained from ClassyFire.²¹ (c) Distribution of the QEDw scores for the 228 drug-like secondary metabolites that satisfy all the 6 drug-likeness scoring schemes.

MeFSAT database. In the CSNs, the nodes representing secondary metabolites are colored in green if they are similar to any of the FDA approved drugs, else they are colored in pink (Methods; Fig. 6a and ESI Fig. S2†). Moreover, the thicknesses of edges in CSNs reflect the chemical similarity between pairs of secondary metabolites connected by them.

Specifically, we find that 82 of 1830 secondary metabolites and 6 of 228 drug-like secondary metabolites in MeFSAT database have high similarity to at least one of the FDA approved drugs. The two CSNs contain multiple disconnected clusters and several isolated nodes which underscore the rich structural diversity in the chemical space of secondary metabolites in MeFSAT database (Fig. 6a and ESI Fig. S2†). Specifically, the



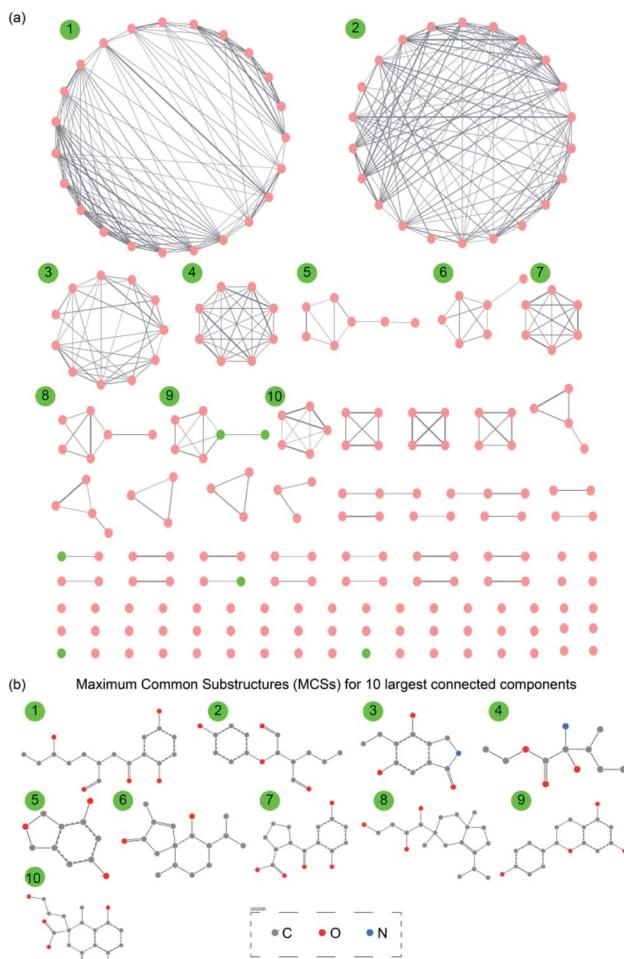


Fig. 6 (a) Chemical similarity network (CSN) of 228 drug-like secondary metabolites in MeFSAT database. Here, the node color is green if the corresponding secondary metabolite is similar to any of the FDA approved drugs else the node color is pink. Edge thickness is proportional to the computed structural similarity between pairs of secondary metabolites. (b) Maximum Common Substructures (MCSs) for 10 largest connected components in the CSN of drug-like secondary metabolites. For this figure, the visualization of the SMARTS was generated using SMARTSview.⁵⁹

CSN of 1830 secondary metabolites has 335 connected components of which 206 are isolated nodes (ESI Fig. S2†). Similarly, the CSN of 228 drug-like secondary metabolites has 94 connected components of which 55 are isolated nodes (Fig. 6a). Moreover, graph density which is a measure of the fraction of all possible edges that are realized in the network, is found to be 0.01 and 0.02, respectively, for the CSNs of 1830 secondary metabolites and 228 drug-like secondary metabolites, respectively. This overly sparse nature of the two CSNs further underscores the structural diversity of the chemical space of secondary metabolites in MeFSAT database.

Finally, we have also computed the maximum common substructures (MCSs) for the 10 largest connected components within the CSN of 228 drug-like secondary metabolites (Methods; Fig. 6b and ESI Table S10†). The computed MCSs display the unique structural motifs underlying the 10 largest

connected components in the CSN of drug-like secondary metabolites (Fig. 6b). In future, it will be worthwhile to further explore the scaffold diversity⁶⁵ of these drug-like fungal secondary metabolites in MeFSAT.

Conclusions

There is immense interest in tapping the diverse chemical space of fungal secondary metabolites for drug discovery. For centuries, medicinal fungi including mushrooms have been used in many systems of traditional medicine to treat human ailments. Since the therapeutic action of medicinal fungi is likely due to their novel secondary metabolites, a curated database compiling information on secondary metabolites and therapeutic uses of medicinal fungi will be a valuable resource for computer-aided drug discovery. Therefore, we have built the first dedicated resource MeFSAT compiling information on secondary metabolites and therapeutic uses of medicinal fungi from published literature. MeFSAT compiles manually curated information on 184 medicinal fungi, 1830 secondary metabolites and 149 therapeutic uses from published literature. For the non-redundant library of 1830 secondary metabolites, we have compiled information on their chemical structure, physicochemical properties, drug-likeness properties, predicted ADMET properties, molecular descriptors, and predicted human target proteins. MeFSAT database also compiles taxonomic information, genome sequencing status, system of traditional medicine and therapeutic uses of medicinal fungi.

After building the MeFSAT database, we have compared the distributions of physicochemical properties for the 1830 secondary metabolites with those for three other small molecule collections. Based on this comparative analysis, we find that the stereochemical complexity and shape complexity of secondary metabolites in MeFSAT are similar to those for other natural product libraries. Based on the previous work by Clemons *et al.*,⁵⁵ one can also extrapolate that the secondary metabolites of medicinal fungi are likely to be specific binders of target proteins. Using multiple scoring schemes, we have also filtered a subset of 228 drug-like secondary metabolites within the MeFSAT database. Lastly, the construction and analysis of chemical similarity networks of secondary metabolites underscores the high structural diversity of the associated chemical space. In conclusion, MeFSAT is a curated resource of fungal natural products which will facilitate traditional knowledge based drug discovery.

Author contributions

R. P. Vivek-Ananth: Data curation, formal analysis, software, visualization, writing – original draft. Ajaya Kumar Sahoo: Data curation, formal analysis, software, visualization, writing – original draft. Kavyaa Kumaravel: Data curation, writing – original draft. Karthikeyan Mohanraj: Software, visualization. Areejit Samal: Conceptualization, formal analysis, supervision, writing – original draft, writing – review & editing.



Conflicts of interest

The authors declare that they have no known conflicts of interest.

Acknowledgements

We thank B. Raveendra Reddy for computational support. AS acknowledges support from a Ramanujan fellowship (SB/S2/RJN-006/2014) from the Science and Engineering Research Board (SERB) India, the Department of Atomic Energy (DAE) India, and a Max Planck Partner Group in Mathematical Biology from the Max Planck Society Germany. The funders have no role in study design, data collection, data analysis, manuscript preparation or decision to publish.

References

- 1 J. Macheleidt, D. J. Mattern, J. Fischer, T. Netzker, J. Weber, V. Schroechk, V. Valiante and A. A. Brakhage, *Annu. Rev. Genet.*, 2016, **50**, 371–392.
- 2 A. K. Singh, H. K. Rana and A. K. Pandey, in *Recent Advancement in White Biotechnology Through Fungi*, ed. A. N. Yadav, S. Singh, S. Mishra and A. Gupta, Springer International Publishing, Cham, 2019, pp. 229–248.
- 3 P.-F. Hu, J. Huang, L. Chen, Z. Ding, L. Liu, I. Molnár and B.-B. Zhang, *J. Agric. Food Chem.*, 2020, **68**, 3995–4004.
- 4 N. P. Keller, *Nat. Rev. Microbiol.*, 2019, **17**, 167–180.
- 5 O. Mosunova, J. C. Navarro-Muñoz and J. Collemare, *Reference Module in Life Sciences*, Elsevier, 2020, p. B9780128096338210000.
- 6 D. J. Newman and G. M. Cragg, *J. Nat. Prod.*, 2016, **79**, 629–661.
- 7 M. Dunkel, M. Fullbeck, S. Neumann and R. Preissner, *Nucleic Acids Res.*, 2006, **34**, D678–D683.
- 8 C. Y.-C. Chen, *PLoS One*, 2011, **6**, e15939.
- 9 F. Ntie-Kang, D. Zofou, S. B. Babiaka, R. Meudom, M. Scharfe, L. L. Lifongo, J. A. Mbah, L. M. Mbaze, W. Sippl and S. M. N. Efange, *PLoS One*, 2013, **8**, e78085.
- 10 T. Xie, S. Song, S. Li, L. Ouyang, L. Xia and J. Huang, *Cell Proliferation*, 2015, **48**, 398–404.
- 11 R. Zhang, S. Yu, H. Bai and K. Ning, *Sci. Rep.*, 2017, **7**, 2821.
- 12 S. Pathania, S. M. Ramakrishnan and G. Bagler, *Database*, 2015, **2015**, bav075.
- 13 L. Huang, D. Xie, Y. Yu, H. Liu, Y. Shi, T. Shi and C. Wen, *Nucleic Acids Res.*, 2018, **46**, D1117–D1120.
- 14 K. Mohanraj, B. S. Karthikeyan, R. P. Vivek-Ananth, R. P. B. Chand, S. R. Aparna, P. Mangalapandi and A. Samal, *Sci. Rep.*, 2018, **8**, 4329.
- 15 J. A. van Santen, G. Jacob, A. L. Singh, V. Aniebok, M. J. Balunas, D. Bunko, F. C. Neto, L. Castaño-Espriu, C. Chang, T. N. Clark, J. L. Cleary Little, D. A. Delgadillo, P. C. Dorresteijn, K. R. Duncan, J. M. Egan, M. M. Gale, F. P. J. Haeckl, A. Hua, A. H. Hughes, D. Iskakova, A. Khadilkar, J.-H. Lee, S. Lee, N. LeGrow, D. Y. Liu, J. M. Macho, C. S. McCaughey, M. H. Medema, R. P. Neupane, T. J. O'Donnell, J. S. Paula, L. M. Sanchez, A. F. Shaikh, S. Soldatou, B. R. Terlouw, T. A. Tran, M. Valentine, J. J. J. van der Hooft, D. A. Vo, M. Wang, D. Wilson, K. E. Zink and R. G. Linington, *ACS Cent. Sci.*, 2019, **5**, 1824–1833.
- 16 X. Zeng, P. Zhang, Y. Wang, C. Qin, S. Chen, W. He, L. Tao, Y. Tan, D. Gao, B. Wang, Z. Chen, W. Chen, Y. Y. Jiang and Y. Z. Chen, *Nucleic Acids Res.*, 2019, **47**, D1118–D1127.
- 17 C. Hobbs, *Medicinal mushrooms: an exploration of tradition, healing, and culture*, Botanica Press, Summertown, TN, 2002.
- 18 M. Powell, *Medicinal mushrooms. The Essential Guide*, Mycology Press, Oxfordshire, UK, 2013.
- 19 J. Meuninck, *Basic illustrated edible and medicinal mushrooms*, FalconGuides, Guilford, Connecticut, 2015.
- 20 *Medicinal plants and fungi: recent advances in research and development*, ed. D. C. Agrawal, H.-S. Tsay, L.-F. Shyur, Y.-C. Wu and S.-Y. Wang, Springer, Singapore, 1st edn, 2017.
- 21 *Medicinal mushrooms: recent progress in research and development*, ed. D. C. Agrawal and M. Dhanasekaran, Springer, Singapore, 1st edn, 2019.
- 22 F.-F. Wang, C. Shi, Y. Yang, Y. Fang, L. Sheng and N. Li, *Biomed. Pharmacother.*, 2018, **102**, 18–25.
- 23 M. E. Valverde, T. Hernández-Pérez and O. Paredes-López, *Int. J. Microbiol.*, 2015, **2015**, 1–14.
- 24 B.-Z. Zaidman, M. Yassin, J. Mahajna and S. P. Wasser, *Appl. Microbiol. Biotechnol.*, 2005, **67**, 453–468.
- 25 *Biosynthesis and molecular genetics of fungal secondary metabolites*, ed. J. F. Martín, C. García-Estrada and S. Zeilinger, Springer, New York, 2014.
- 26 E. W. Sayers, J. Beck, J. R. Brister, E. E. Bolton, K. Canese, D. C. Comeau, K. Funk, A. Ketter, S. Kim, A. Kimchi, P. A. Kitts, A. Kuznetsov, S. Lathrop, Z. Lu, K. McGarvey, T. L. Madden, T. D. Murphy, N. O'Leary, L. Phan, V. A. Schneider, F. Thibaud-Nissen, B. W. Trawick, K. D. Pruitt and J. Ostell, *Nucleic Acids Res.*, 2020, **48**, D9–D16.
- 27 *Species 2000 & ITIS Catalogue of Life, 2019 Annual Checklist*, ed. Y. Roskov, G. Ower, T. Orrell, D. Nicolson, N. Bailly, P. M. Kirk, T. Bourgoin, R. E. DeWalt, W. Decock, E. V. Nieuwkerken, J. Zarucchi and L. Penev, Leiden, the Netherlands, 2019, digital resource at www.catalogueoflife.org/annual-checklist/2019.
- 28 V. Robert, D. Vu, A. B. H. Amor, N. van de Wiele, C. Brouwer, B. Jabas, S. Szoke, A. Dridi, M. Triki, S. Ben Daoud, O. Chouchen, L. Vaas, A. de Cock, J. A. Stalpers, D. Stalpers, G. J. M. Verkley, M. Groenewald, F. B. Dos Santos, G. Stegehuis, W. Li, L. Wu, R. Zhang, J. Ma, M. Zhou, S. P. Gorjón, L. Eurwilaichitr, S. Ingsriswang, K. Hansen, C. Schoch, B. Robbertse, L. Irinyi, W. Meyer, G. Cardinali, D. L. Hawksworth, J. W. Taylor and P. W. Crous, *IMA Fungus*, 2013, **4**, 371–379.
- 29 J. D. Lawrey and P. Diederich, *Bryologis*, 2003, **106**, 80–120.
- 30 *Fungal metabolites*, ed. J.-M. Merillon and K. G. Ramawat, Springer International Publishing, Switzerland, 2017.
- 31 S. Kim, P. A. Thiessen, E. E. Bolton, J. Chen, G. Fu, A. Gindulyte, L. Han, J. He, S. He, B. A. Shoemaker, J. Wang, B. Yu, J. Zhang and S. H. Bryant, *Nucleic Acids Res.*, 2016, **44**, D1202–D1213.

32 H. E. Pence and A. Williams, *J. Chem. Educ.*, 2010, **87**, 1123–1124.

33 T. T. Tanimoto, *An Elementary Mathematical theory of Classification and Prediction*, IBM, 1957.

34 RDKit, *Open-source cheminformatics*, <http://www.rdkit.org>.

35 S. Riniker and G. A. Landrum, *J. Chem. Inf. Model.*, 2015, **55**, 2562–2574.

36 T. A. Halgren, *J. Comput. Chem.*, 1996, **17**, 490–519.

37 N. M. O'Boyle, M. Banck, C. A. James, C. Morley, T. Vandermeersch and G. R. Hutchison, *J. Cheminf.*, 2011, **3**, 33.

38 Y. Djoumbou Feunang, R. Eisner, C. Knox, L. Chepelev, J. Hastings, G. Owen, E. Fahy, C. Steinbeck, S. Subramanian, E. Bolton, R. Greiner and D. S. Wishart, *J. Cheminf.*, 2016, **8**, 61.

39 A. Daina, O. Michielin and V. Zoete, *Sci. Rep.*, 2017, **7**, 42717.

40 C. A. Lipinski, F. Lombardo, B. W. Dominy and P. J. Feeney, *Adv. Drug Delivery Rev.*, 2001, **46**, 3–26.

41 A. K. Ghose, V. N. Viswanadhan and J. J. Wendoloski, *J. Comb. Chem.*, 1999, **1**, 55–68.

42 D. F. Veber, S. R. Johnson, H.-Y. Cheng, B. R. Smith, K. W. Ward and K. D. Kopple, *J. Med. Chem.*, 2002, **45**, 2615–2623.

43 W. J. Egan, K. M. Merz and J. J. Baldwin, *J. Med. Chem.*, 2000, **43**, 3867–3877.

44 J. D. Hughes, J. Blagg, D. A. Price, S. Bailey, G. A. Decrescenzo, R. V. Devraj, E. Ellsworth, Y. M. Fobian, M. E. Gibbs, R. W. Gilles, N. Greene, E. Huang, T. Krieger-Burke, J. Loesel, T. Wager, L. Whiteley and Y. Zhang, *Bioorg. Med. Chem. Lett.*, 2008, **18**, 4872–4875.

45 M. P. Gleeson, *J. Med. Chem.*, 2008, **51**, 817–834.

46 S. J. Teague, A. M. Davis, P. D. Leeson and T. Oprea, *Angew. Chem., Int. Ed.*, 1999, **38**, 3743–3748.

47 G. R. Bickerton, G. V. Paolini, J. Besnard, S. Muresan and A. L. Hopkins, *Nat. Chem.*, 2012, **4**, 90–98.

48 C. W. Yap, *J. Comput. Chem.*, 2011, **32**, 1466–1474.

49 I. V. Grigoriev, H. Nordberg, I. Shabalov, A. Aerts, M. Cantor, D. Goodstein, A. Kuo, S. Minovitsky, R. Nikitin, R. A. Ohm, R. Otiilar, A. Poliakov, I. Ratnere, R. Riley, T. Smirnova, D. Rokhsar and I. Dubchak, *Nucleic Acids Res.*, 2012, **40**, D26–D32.

50 F. B. Rogers, *Bull. Med. Libr. Assoc.*, 1963, **51**, 114–116.

51 L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng and W. A. Kibbe, *Nucleic Acids Res.*, 2012, **40**, D940–D946.

52 Medicode (Firm), *ICD-9-CM: International classification of diseases, 9th revision, clinical modification*, Medicode, Salt Lake City, Utah, 5th edn, 1997.

53 D. Szklarczyk, A. Santos, C. von Mering, L. J. Jensen, P. Bork and M. Kuhn, *Nucleic Acids Res.*, 2016, **44**, D380–D384.

54 E. A. Bruford, M. J. Lush, M. W. Wright, T. P. Sneddon, S. Povey and E. Birney, *Nucleic Acids Res.*, 2008, **36**, D445–D448.

55 P. A. Clemons, N. E. Bodycombe, H. A. Carrinski, J. A. Wilson, A. F. Shamji, B. K. Wagner, A. N. Koehler and S. L. Schreiber, *Proc. Natl. Acad. Sci. U. S. A.*, 2010, **107**, 18787–18792.

56 D. Bajusz, A. Rácz and K. Héberger, *J. Cheminf.*, 2015, **7**, 20.

57 D. Rogers and M. Hahn, *J. Chem. Inf. Model.*, 2010, **50**, 742–754.

58 D. S. Wishart, Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Lynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox and M. Wilson, *Nucleic Acids Res.*, 2018, **46**, D1074–D1082.

59 S. Jasial, Y. Hu, M. Vogt and J. Bajorath, *F1000Research*, 2016, **5**, 591.

60 M. Franz, C. T. Lopes, G. Huck, Y. Dong, O. Sumer and G. D. Bader, *Bioinformatics*, 2016, **32**, 309–311.

61 Y. Cao, T. Jiang and T. Girke, *Bioinformatics*, 2008, **24**, i366–i374.

62 K. Schomburg, H.-C. Ehrlich, K. Stierand and M. Rarey, *J. Chem. Inf. Model.*, 2010, **50**, 1529–1535.

63 C.-Y. Jia, J.-Y. Li, G.-F. Hao and G.-F. Yang, *Drug Discovery Today*, 2020, **25**, 248–258.

64 F. E. Koehn, *MedChemComm*, 2012, **3**, 854.

65 M. González-Medina, J. R. Owen, T. El-Elimat, C. J. Pearce, N. H. Oberlies, M. Figueroa and J. L. Medina-Franco, *Front. Pharmacol.*, 2017, **8**, 180.

