

**Compilation, curation and exploration
of natural product spaces to enable
traditional knowledge based
drug discovery**

By

Vivek Ananth R.P.

LIFE10201604001

**The Institute of Mathematical Sciences
Chennai**

*A thesis submitted to the
Board of Studies in Life Sciences
In partial fulfillment of requirements
for the Degree of*

DOCTOR OF PHILOSOPHY

of

HOMI BHABHA NATIONAL INSTITUTE

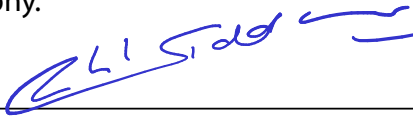


July 2022


Homi Bhabha National Institute

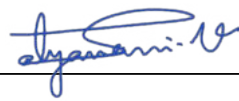
Recommendations of the Viva Voce Committee

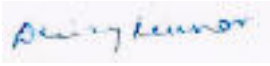
As members of the Viva Voce Committee, we certify that we have read the dissertation prepared by Wivek Ananth R.P. entitled: "Compilation, curation and exploration of natural product spaces to enable traditional knowledge based drug discovery" and recommend that it may be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.


_____ Date: 22/11/2022
Chair - Prof. Rahul Siddharthan


_____ Date: 22/11/2022
Supervisor/Convener - Prof. Areejit Samal


_____ Date: 22/11/2022
Member 1 - Prof. Sitabhra Sinha


_____ Date: 22/11/2022
Member 2 - Prof. Satyavani Vemparala


_____ Date: 22/11/2022
Member 3 - Prof. Dhiraj Kumar


_____ Date: 22/11/2022
External Examiner - Prof. Sudip Kundu

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to HBNI.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it may be accepted as fulfilling the dissertation requirement.

Date: 22/11/2022

Place: CHENNAI


Supervisor

Statement by Author

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at Homi Bhabha National Institute (HBNI) and is deposited in the Library to be made available to borrowers under rules of the HBNI.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgement of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the Competent Authority of HBNI when in his or her judgement the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.



Vivek Ananth R.P.

Declaration

I, hereby declare that the investigation presented in this thesis has been carried out by me. The work is original and has not been submitted earlier as a whole or in part for a degree or diploma at this or any other Institution or University.

A handwritten signature in blue ink, appearing to read 'Vivek Ananth R.P.', is positioned above the printed name.

Vivek Ananth R.P.

List of Publications arising from the thesis

Journals

Published

1. *IMPPAT: A curated database of Indian Medicinal Plants, Phytochemistry and Therapeutics*, K. Mohanraj[†], B.S. Karthikeyan[†], **R.P. Vivek-Ananth**[†], R.P. Bharath Chand, S.R. Aparna, P. Mangalapandi and A. Samal^{*}, *Scientific Reports*, 8: 4329 (2018). <https://doi.org/10.1038/s41598-018-22631-z>
2. *In Silico Identification of Potential Natural Product Inhibitors of Human Proteases Key to SARS-CoV-2 Infection*, **R.P. Vivek-Ananth**, A. Rana, N. Rajan, H.S. Biswal^{*} and A. Samal^{*}, *Molecules*, 25(17): 3822 (2020). <https://doi.org/10.3390/molecules25173822>
3. *MeFSAT: A curated natural product database specific to secondary metabolites of medicinal fungi*, **R.P. Vivek-Ananth**[†], A.K. Sahoo[†], K. Kumaravel[†], K. Mohanraj and A. Samal^{*}, *RSC Advances*, 11: 2596-2607 (2021). <https://doi.org/10.1039/D0RA10322E>
4. *Potential phytochemical inhibitors of SARS-CoV-2 helicase Nsp13: a molecular docking and dynamic simulation study*, **R.P. Vivek-Ananth**, S. Krishnaswamy^{*} and A. Samal^{*}, *Molecular Diversity*, 26: 429-442 (2022). <https://doi.org/10.1007/s11030-021-10251-1>

Submitted

5. *IMPPAT 2.0: an enhanced and expanded phytochemical atlas of Indian medicinal plants*, **R. P. Vivek-Ananth**, K. Mohanraj, A.K. Sahoo and A. Samal^{*}, *bioRxiv* 2022.06.17.496609 (2022). <https://doi.org/10.1101/2022.06.17.496609>

[[†] Joint-first authors; ^{*} Corresponding author(s)]

List of Publications not included in the thesis

Journals

1. *Comparative systems analysis of the secretome of the opportunistic pathogen Aspergillus fumigatus and other Aspergillus species*. **R.P. Vivek-Ananth**, K. Mohanraj, M. Vandanasree, A. Jhingran, J.P. Craig and A. Samal^{*}, Scientific Reports, 8: 6617 (2018). <https://doi.org/10.1038/s41598-018-25016-4>
2. *Network approach towards understanding the crazing in glassy amorphous polymers*, S. Venkatesan[†], **R.P. Vivek-Ananth**[†], R.P. Sreejith, P. Mangalapandi, A.A. Hassanali^{*} and A. Samal^{*}, Journal of Statistical Mechanics: Theory and Experiment, 043305 (2018). <http://dx.doi.org/10.1088/1742-5468/aab688>
3. *Discrete Ricci curvatures for directed networks*, E. Saucan, R.P. Sreejith, **R.P. Vivek-Ananth**, J. Jost and A. Samal^{*}, Chaos, Solitons & Fractals, 118: 347-360 (2019). <https://doi.org/10.1016/j.chaos.2018.11.031>
4. *A curated knowledgebase on endocrine disrupting chemicals and their biological systems-level perturbations*, B.S. Karthikeyan[†], J. Ravichandran^{†,*}, K. Mohanraj^γ, **R.P. Vivek-Ananth**^γ and A. Samal^{*}, Science of the Total Environment, 692: 281-296 (2019). <https://doi.org/10.1016/j.scitotenv.2019.07.225>
5. *Reprogramming of microRNA expression via E2F1 downregulation promotes Salmonella infection both in infected and bystander cells*, C. Aguilar, S. Costa, C. Maudet, **R.P. Vivek-Ananth**, S. Zaldívar-López, J.J. Garrido, A. Samal, M. Mano and A. Eulalio^{*}, Nature Communications, 12: 3392 (2021). <https://doi.org/10.1038/s41467-021-23593-z>
6. *Virtual screening of phytochemicals from Indian medicinal plants against the endonuclease domain of SFTS virus L polymerase*, **R.P. Vivek-Ananth**, A.K. Sahoo, A. Srivastava^{*} and A. Samal^{*}, RSC Advances, 12: 6234-6247 (2022). <https://doi.org/10.1039/D1RA06702H>

[[†] Joint-first authors; ^γ Joint-second authors; ^{*} Corresponding author(s)]

Copyright

1. *DEDuCT - Database of Endocrine Disrupting Chemicals and their Toxicity Profiles*. Authors: A. Samal, J. Ravichandran, B.S. Karthikeyan, M. Karthikeyan and **R.P. Vivek Ananth**.

Copyright granted to The Institute of Mathematical Sciences by the Copyright office, Government of India, with the Diary Number 16429/2018-CO/L.

2. *IMPPATx: A Curated Expanded Database of Indian Medicinal Plants, Phytochemicals, And Therapeutic Uses*. Authors: A. Samal, **R.P. Vivek Ananth** and M. Karthikeyan.

Copyright granted to The Institute of Mathematical Sciences by the Copyright office, Government of India, with the Diary Number 11598/2019-CO/L.

Oral or Poster presentations

1. Poster presentation titled *Protein secretion machinery in a model filamentous fungus* at the symposium on Systems, Synthetic & Chemical Biology held at the Bose Institute, Kolkata from December 5-7, 2017.
2. Poster presentation titled *Prediction and analysis of the secretome of an opportunistic fungal pathogen* at the 17th International Conference on Bioinformatics (InCoB 2018) held at the Jawaharlal Nehru University (JNU), New Delhi from September 26-28, 2018.
3. Poster presentation titled *Systems modelling of protein secretion system in model filamentous fungus* at the 87th Annual meeting and conference of Society of Biological Chemists (India) (SBCI-2018) held at the Manipal Academy of Higher Education, Manipal from November 25-27, 2018.
4. Poster presentation titled *Prediction and analysis of the secretome of an opportunistic fungal pathogen* at the 11th International Conference on Biology of Yeasts and Filamentous Fungi (ICBYFF-2019) held at the University of Hyderabad, Hyderabad from November 27-29, 2019.

5. Oral presentation titled *In silico identification of potential natural product inhibitors of human proteases key to SARS-CoV-2 infection* at the Young Scientists' conference of the India International Science Festival (IISF-2020) held from December 22-24, 2020.

Research visits and seminars

1. Seminar titled *Systems modeling of protein secretion system in model filamentous fungus* at the Institute Seminar Days 2017, held at The Institute of Mathematical Sciences (IMSc), Chennai on March 15, 2017.
2. Seminar titled *Prediction and analysis of the secretome of an opportunistic fungal pathogen* at the Institute Seminar Days 2018, held at The Institute of Mathematical Sciences (IMSc), Chennai on April 9, 2018.
3. Seminar titled *Prediction and analysis of the secretome of an opportunistic fungal pathogen & Exploration of the phytochemical space of Indian medicinal plants* at the Quantitative Life Sciences section, The Abdus Salam International Centre for Theoretical Physics (ICTP), Trieste, Italy on September 10, 2018.



Vivek Ananth R.P.

This thesis is dedicated
to my parents

R.Palanikumar & M.Vijayalakshmi

for their love, support and encouragement

Acknowledgements

First of all, I am deeply indebted to my thesis supervisor Prof. Areejit Samal for guiding me through these years of PhD. He has not just guided me on the research projects that I undertook during the course of my PhD, but has also imparted a great deal of practical details on how to thrive in the highly competitive research atmosphere. Through the course of several years of working with him, I have imbibed his passion towards science and his unrelenting commitment to finish the projects even amidst several odds. Any number of sentences I write thanking him will not be able to convey my gratitude to him. I can only feel grateful for having him as my thesis supervisor.

I would like to extend my sincere thanks to all my co-authors, specifically M. Karthikeyan, Dr. B.S. Karthikeyan, Dr. R. Janani, Ajaya Kumar Sahoo, K. Kavya, R.P. Sreejith, M. Vandanasree, Dr. V. Sudarkodi, Abhijit Rana and P. Mangalapandi for their valuable scientific contributions. I am also grateful to Prof. Vinay K. Nandicoori, Prof. Dhiraj Kumar, Prof. Jürgen Jost, Prof. Ana Eualio, Prof. Miguel Mano, Prof. S. Krishnaswamy, Prof. Ashutosh Srivastava, Prof. Himansu Sekhar Biswal, Prof. Emil Saucan, Prof. Ali Hassanali, Dr. James. P. Craig and Dr. Anupam Jhingran for collaborating with us on several research projects. I am also thankful to Prof. Sanjay Jain, Prof. N. Sukumar and Prof. Matteo Marsili for their insights and discussions. I would like to recognize Pavithra Elumalai, Yasharth Yadav, Subbaroyan Ajay, Dr. Pinaki Saha, D. Gokul Balaji, A. Priya Dharshini, R. Nithin, G. Rajesh, R.P. Bharath Chand, S.R. Aparna and Ashreya Jayaram for their help and discussions. I would also like to thank all the interns and project assistant who have directly or indirectly contributed to the research projects. I also thank the system administrators B. Raveendra Reddy and Dr. G. Subramoniam, G. Srinivasan, Imran Khan and the computer committee of IMSc for their technical support. Lastly, I would like to acknowledge the support by the library staff of IMSc in procuring several books which were digitized to built IMPPAT.

I thank all my doctoral committee members for their constructive comments and feed-

back during the doctoral committee meetings. I would like to thank Prof. Gautam Menon, Prof. Sitabhra Sinha, Prof. Rahul Siddharthan, Prof. Satyavani Vemparala, Prof. S. Krishnaswamy, Prof. Vasudharani Devanathan, Dr. Nivedita Chatterjee, Prof. Amritanshu Prasad and other visiting scientists who taught various subjects during my PhD coursework period. I also thank the administrative, canteen, civil maintenance, electrical, gardening, hostel, house keeping, security staffs and all other permanent and contractual staffs of IMSc for their tireless work in smooth functioning of the institute and thus facilitating our research.

Words cannot express my gratitude to my parents R. Palanikumar and M. Vijayalakshmi for their love and encouragement. They always ensured me and my sister get the best opportunities to excel in academics and extra-curricular activities. I feel overwhelmed to know they will be the happiest people as I complete my research studies in fulfillment for the PhD degree. I cannot miss thanking my sister R.P. Suganya who has been always around supporting me all these years. My heartfelt thanks also goes to my paternal and maternal grandparents, relatives, cousins, nieces and nephews. I am specifically grateful to my late maternal grandmother Murugaiyan Mahalakshmi for all her love and care. I am thankful to all my teachers who taught me right from school to college. Specifically, I would like to thank my teachers Prof. S. Lakshmana Prabu, Prof. Jayaraman Valadi, Prof. Ashutosh Singh, Prof. Pawan Dhar. I also thank Prof. Dasaradhi Palakodeti who guided my Master's thesis research work. It will be injustice if I miss thanking my friends, especially those who kept checking on me even though I was absent in all their major life events. I thank Arun Krishnan, Monish Mohankumar, Magesh Kalaiselvan, Vijay Vignesh, B. Kiran Sabarish, M. Dinesh Pandian, T. Venkatachalapathy, P. Suganya, Dr. Shubhra Agarwal, Dr. Vijeta Sharma, Bhumika Dubay, Dr. Priyamvad Srivastav, Mariya Plackattu and many more I missed here for being around and supporting me through the years. Lastly, I would like to thank Gurudev Sri Sri Ravishankar for being my guiding light and helping me sail through testing times.

Vivek Ananth R.P.

Contents

List of Figures	i
List of Tables	v
Abstract	v
1 Introduction	1
1.1 Motivation	1
1.2 Phytochemical space of Indian medicinal plants	6
1.3 Secondary metabolite space of medicinal fungi	8
1.4 Cheminformatics based analysis of natural product spaces	9
1.5 Thesis organization	11
2 IMPPAT: A curated database of <u>I</u>ndian <u>M</u>edicinal <u>P</u>lants, <u>P</u>hytochemistry <u>A</u>nd <u>T</u>herapeutics	15
2.1 Workflow for construction of IMPPAT 1.0	16
2.1.1 Curated list of Indian medicinal plants	16
2.1.2 Phytochemical composition of Indian medicinal plants	17
2.1.3 Annotation, curation and filtering of identified phytochemicals	20
2.1.4 Therapeutic uses of Indian medicinal plants	21
2.1.5 Traditional formulations of Indian medicinal plants	22
2.2 IMPPAT 2.0: an enhanced and expanded phytochemical atlas of Indian medicinal plants	23
2.2.1 Increase in coverage of Indian medicinal plants	23

2.2.2	Information at the level of plant parts	27
2.2.3	Increase in coverage of phytochemicals	27
2.2.4	Enhanced annotation to enable exploration of the phytochemical space	29
2.2.5	Increase in coverage of therapeutic uses	33
2.2.6	Increase in coverage of traditional medicinal formulations	34
2.3	Web design and data accessibility	35
2.4	Discussion	40
3	Exploration of the phytochemical space of Indian medicinal plants	45
3.1	Molecular complexity comparison with other collections of small molecules	46
3.2	Molecular scaffold based structural diversity	50
3.3	Drug-like phytochemical space	54
3.4	Comparison with the phytochemical space of Chinese medicinal plants . .	58
3.5	Discussion	60
4	Compilation, curation and exploration of a chemical atlas of secondary metabolites from medicinal fungi	65
4.1	Workflow for the compilation and curation of MeFSAT database	66
4.1.1	Compilation of curated list of medicinal fungi	66
4.1.2	Compilation of the secondary metabolites of medicinal fungi	68
4.1.3	Curated <i>in silico</i> library of secondary metabolites of medicinal fungi	68
4.1.4	Annotation of secondary metabolites of medicinal fungi	69
4.1.5	Genome sequences of medicinal fungi	70
4.1.6	Compilation and curation of therapeutic uses of medicinal fungi	71
4.1.7	Predicted human target proteins of secondary metabolites	71
4.2	Web-interface of MeFSAT	73
4.3	Exploration of the curated information on medicinal fungi, their sec- ondary metabolites and therapeutic uses	75

4.4	Comparison of the molecular complexity of secondary metabolites in MeFSAT with other small molecule collections	77
4.5	Drug-like secondary metabolites of medicinal fungi	79
4.6	Chemical similarity networks of secondary metabolites	81
4.7	Discussion	86
5	<i>In silico</i> identification of potential anti-COVID drugs from phytochemicals of Indian medicinal plants	89
5.1	TMPRSS2 and cathepsin L: key human proteases for host cell entry of SARS-CoV-2	91
5.2	SARS-CoV-2 helicase Nsp13	94
5.3	Methods	94
5.3.1	Preparation of ligand library of phytochemicals	94
5.3.2	Molecular docking of the phytochemicals to the target proteins	95
5.3.3	Identification of protein-ligand interactions	96
5.3.4	Molecular Dynamics simulations	97
5.4	Virtual screening for key host factors in SARS-CoV-2 infection	99
5.4.1	Potential Phytochemical Inhibitors of TMPRSS2	103
5.4.2	Potential Phytochemical Inhibitors of Cathepsin L	108
5.5	Virtual screening for SARS-Cov-2 Nsp13	114
5.5.1	Potential Phytochemical Inhibitors of SARS-CoV-2 Nsp13	117
5.6	Discussion	121
6	Summary and future outlook	127
6.1	Summary	128
6.2	Future outlook	132
A	Analysis of top inhibitors predicted for key host factors in SARS-CoV-2 infection	135
A.1	Reference inhibitors of TMPRSS2 and Cathepsin L	135

A.2	Molecular Dynamics simulation of top inhibitors	138
A.3	MM-PBSA binding energy of top inhibitors	144
B	Analysis of top inhibitors predicted for SARS-CoV-2 Nsp13	147
B.1	Comparison with ligands co-crystallized with Nsp13	147
B.2	Molecular Dynamics simulation of top inhibitors	148
B.3	MM-PBSA binding energy of top inhibitors	154
	References	156

List of Figures

1.1	Natural product space, a biologically relevant subspace for drug discovery	2
1.2	Potential importance of digitizing traditional knowledge associated with Indian medicinal plants for natural product based drug discovery	4
2.1	Schematic overview of the IMPPAT 1.0 database construction pipeline . .	19
2.2	Schematic overview of the important features including enhancements and expansion realized in IMPPAT 2.0	24
2.3	Coverage of Indian medicinal plants in IMPPAT 2.0	26
2.4	Basic statistics and distribution of the physicochemical properties for phytochemicals in IMPPAT 2.0	30
2.5	Chemical classification, biosynthetic pathways and natural product likeness of phytochemicals in IMPPAT 2.0	32
2.6	Web-interface of the IMPPAT 1.0 database	37
2.7	Web-interface of the IMPPAT 2.0 database	38
3.1	Comparison of the molecular complexity of IMPPAT 1.0 with other chemical libraries	47
3.2	Comparison of the molecular complexity of IMPPAT 2.0 with other chemical libraries	48
3.3	Analysis of the scaffold diversity of phytochemicals in IMPPAT 2.0 with seven other natural product libraries, approved drugs, and organic compounds from PubChem	51
3.4	Molecular cloud visualization of the top scaffolds at G/N/B level present in phytochemicals of IMPPAT 2.0	53

3.5	Drug-likeness analysis of phytochemicals in IMPPAT 2.0	57
3.6	Chemical similarity network of the 1335 drug-like phytochemicals in IMPPAT 2.0	59
3.7	Comparison of the phytochemical space of Indian medicinal plants and Chinese medicinal plants	61
4.1	Schematic overview of the workflow to construct the MeFSAT database .	67
4.2	Web-interface of the MeFSAT database	73
4.3	Basic statistics for medicinal fungi, their secondary metabolites and ther- apeutic uses in MeFSAT database	76
4.4	Comparison of the stereochemical complexity, shape complexity and physicochemical properties of secondary metabolites in MeFSAT with other small molecule collections	80
4.5	Drug-likeness analysis of the secondary metabolites in MeFSAT database	82
4.6	Histogram showing the number of secondary metabolites in MeFSAT which satisfy at least 1, at least 2, at least 3, at least 4, at least 5 and all 6 of the drug-likeness scoring schemes evaluated here	83
4.7	Chemical similarity network (CSN) of 1830 secondary metabolites in MeFSAT database	84
4.8	Chemical similarity network (CSN) of 228 drug-like secondary metabo- lites in MeFSAT database	85
5.1	Cartoon representation of the homology model structure of TMPRSS2 . .	93
5.2	Cartoon representation of the crystal structure of human cathepsin L . . .	93
5.3	Cartoon representation of the prepared crystal structure of SARS-CoV-2 helicase Nsp13	95
5.4	Geometric criteria for the identification of protein-ligand interactions . . .	98
5.5	Workflow to identify potential phytochemical inhibitors of human pro- teases TMPRSS2 and cathepsin L	99

5.6	Molecular structure and chemical name of the top 9 phytochemical inhibitors (compounds T1–T9) of TMPRSS2	104
5.7	Cartoon representation of the protein-ligand interactions of the phytochemical inhibitors of TMPRSS2	106
5.8	Molecular structures of the top 9 phytochemical inhibitors of cathepsin L .	109
5.9	Cartoon representation of the protein-ligand interactions of the phytochemical inhibitors of cathepsin L	112
5.10	Workflow for the identification of potential phytochemical inhibitors of SARS-CoV-2 helicase Nsp13	115
5.11	Molecular structure and chemical name of the top 10 phytochemical inhibitors (compounds H1–H10) of SARS-CoV-2 Nsp13	119
5.12	Cartoon representation of the hydrogen bond interactions in the best-docked pose of the top 10 potential phytochemical inhibitors of SARS-CoV-2 helicase Nsp13	121
6.1	Summary of the research on compilation, curation and exploration of natural product spaces reported in this thesis	128
A.1	Cartoon representation of the protein-ligand interactions of the known inhibitors of TMPRSS2 and cathepsin L	137
A.2	Superimposition of the docked pose of GH4 with cathepsin L and the pose of GH4 in the co-crystallized structure with cathepsin L	138
A.3	Radius of gyration, RMSD, RMSF and distance from key binding site residue derived from MD simulation of TMPRSS2-ligand complex	139
A.4	Radius of gyration, RMSD, RMSF and distance from key binding site residue derived from MD simulation of cathepsin L-ligand complex . . .	140
A.5	Radius of gyration, RMSD and RMSF from MD simulation of uncomplexed TMPRSS2 free protein and cathepsin L free protein	142

A.6	Superimposition of the protein-ligand complex snapshots from MD simulation trajectories	143
B.1	Binding mode of the top inhibitors of SARS-CoV-2 Nsp13 and the ligands from PanDDA co-crystallized structures	149
B.2	Radius of gyration, RMSD and RMSF derived from MD simulation of Nsp13-ligand complex	150
B.3	Radius of gyration, RMSD and RMSF derived from MD simulation of Nsp13 uncomplexed protein	151
B.4	Superimposition of the protein-ligand complex snapshots from MD simulation trajectories	152
B.5	Distance from key binding site residue derived from MD simulation of Nsp13-ligand complex	153

List of Tables

2.1	Comparison of IMPPAT 1.0 with earlier databases on phytochemical composition of Indian medicinal plants	43
2.2	Comparison of the updated version IMPPAT 2.0 with the previous version 1.0	44
3.1	Scaffold diversity of phytochemicals in IMPPAT 2.0, and comparison with other chemical libraries	63
5.1	Herbal sources of top 9 phytochemical inhibitors of TMPRSS2	124
5.2	Herbal sources of top 9 phytochemical inhibitors of Cathepsin L	125
5.3	Herbal sources of top 10 phytochemical inhibitors of SARS-CoV-2 Nsp13	126
A.1	MM-PBSA based binding energy for the top three inhibitors of TMPRSS2 and cathepsin L	146
B.1	MM-PBSA based binding energy for top five inhibitors of SARS-CoV-2 Nsp13	155

Abstract

Endogenous secondary metabolites produced by living organisms are valuable natural products. Such secondary metabolites are produced primarily by plants, fungi and bacteria. Many natural products have been found to have useful therapeutic activity. By one estimate nearly 34% of the approved drugs are either natural products or natural product derived. Thus, characterization of novel natural product spaces will facilitate identification and development of new drugs. Medicinal plants and medicinal fungi have been used in traditional medicine across the world for treating many human ailments. Therefore, mapping the natural product space of medicinal plants and medicinal fungi will enable effective exploration of the natural product space for drug discovery.

In this thesis, we first focus on the compilation, curation and exploration of the natural product space of Indian medicinal plants. Specifically, we have built a comprehensive database, IMPPAT (version 1.0 and 2.0), on Indian medicinal plants, their phytochemicals and therapeutic uses. IMPPAT provides a FAIR compliant non-redundant *in silico* stereo-aware library of 17967 phytochemicals with 2D and 3D chemical structures. Using cheminformatics based analysis, 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. The compiled information in IMPPAT is accessible at: <https://cb.imsc.res.in/imppat/>.

We have then built a manually curated database MeFSAT dedicated to secondary metabolites and therapeutic uses of medicinal fungi. MeFSAT provides information on 184 medicinal fungi, 1830 secondary metabolites and 149 therapeutic uses obtained from published research articles and books. We have further analyzed the curated secondary metabolite space and have presented the unique features of this natural product space. The compiled information in MeFSAT is accessible at: <https://cb.imsc.res.in/mefsat/>.

The thesis lastly describes a specific biological application of the curated phytochem-

ical library from IMPPAT to identify potential anti-COVID lead molecules. In sum, the manually curated natural product spaces described in this thesis will enrich the known natural product space and will enable traditional knowledge based drug discovery.

Chapter 1

Introduction

1.1 Motivation

Let food be thy medicine and medicine be thy food.

- Hippocrates

Diverse flora and fauna arising from millions of years of evolution inhabit the earth. It is estimated that there are nearly 8.7 million different species of animals, plants and fungi globally, of which only 1.2 million have been characterized [1]. These organisms are distributed across different ecosystems such as marine, rain forests, deserts, etc. Several biotic and abiotic factors of these ecosystems have been considered to be one of the evolutionary forces behind emergence of novel traits (phenotype) in the organisms which inhabit these ecosystems [2, 3]. The metabolome of the organism has been attributed to be the key factor linking the encoded genome to the observed phenotype of the organism [4]. The endogenous metabolome of the organism is defined as the complete collection of small molecules (also called metabolites) naturally produced by the organism [5]. The endogenous metabolome can be further sub-classified as primary and secondary metabolome. The primary metabolites directly influence the growth, development and reproduction of the organism. On the other hand, the secondary metabolites

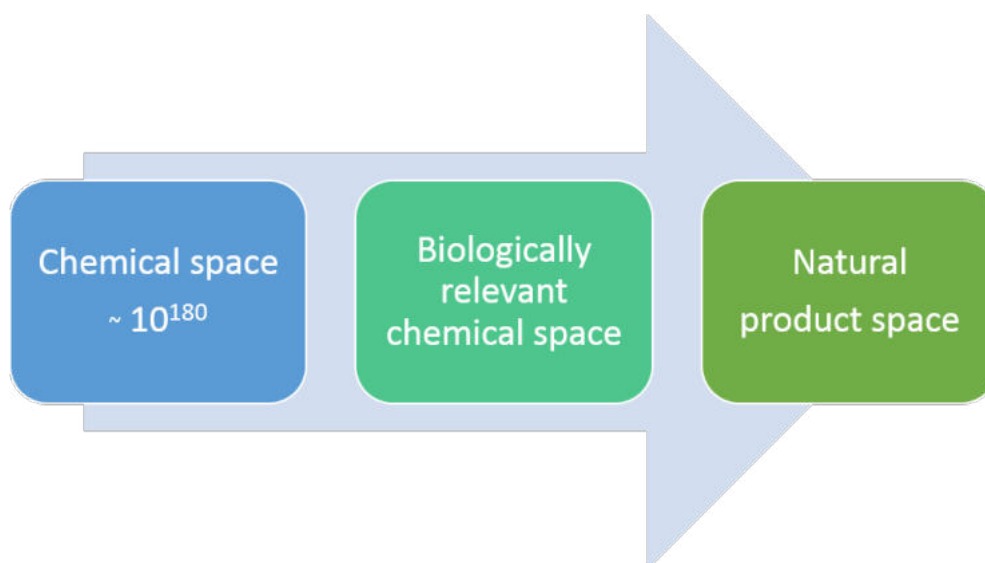


Figure 1.1: Chemical space is vast with a predicted upper limit of approximately 10^{180} molecules [11]. The biologically relevant chemical space is the subspace of interest for drug discovery. Natural products is a subspace of the biologically relevant chemical space curated by nature.

although not directly involved in the growth, development and reproduction of the organism, yet usually have an important ecological function. Secondary metabolites are mainly produced by plants, fungi and bacteria, and secondary metabolites provide these organisms with an evolutionary advantage to survive in diverse ecological niche [6]. For example plant secondary metabolites, also called phytochemicals, play an important role in plant's defence against plant pathogens and herbivory [7,8]. These special metabolites produced naturally by the organisms from different taxonomic groups are collectively referred to as natural products [9]. Apart from being beneficial to the source organism, many natural products have been found to have therapeutic value. Prominent examples of therapeutic natural products include the anti-malarial drug Artemisinin isolated from the plant *Artemisia annua* and anti-cancer drug Paclitaxel isolated from the plant *Taxus brevifolia* [10]. Therefore, natural products continue to be of immense interest to scientist working in the field of drug discovery.

Mapping and exploration of the chemical space of natural products (natural product space) is a crucial step towards the discovery of new therapeutic molecules and better understanding of the biosynthesis of the natural products [12]. Chemical space in gen-

eral encompasses all possible organic chemicals and can be considered similar to the cosmological universe in its vastness [13, 14]. Even limiting to reasonably sized small molecules with less than 30 carbon atoms, it is estimated that there are close to 10^{63} stable molecules, which is astronomically large [15]. Biologically relevant chemical space, which is a subspace of complete chemical space, is still vast, and this is the subspace of interest in research towards the development of new therapeutics. Natural products which are produced by living organisms carry out specific biological functions and provide a selective advantage to the producing organism. The selective pressure experienced by the organism during its course of evolution has led to natural products with diverse molecular scaffolds which selectively modulate the biological targets either directly or indirectly to offer the organism with higher level of fitness [12]. Thus, the natural product space can be considered a biologically relevant chemical space curated by nature [12] (Figure 1.1).

Natural products play an important role in the pharmaceutical industry as new sources of drugs [16–20]. However, recently there has been a decline in the number of marketable drugs derived from natural products [19, 20]. Furthermore, the majority of these drugs fall into already known structural scaffolds as due importance has not been given to unexplored sources of natural products for drug discovery [20]. As a result, lately, there has been significant interest in applying interdisciplinary approaches [21] to explore the novel chemical scaffolds of natural product space for drug discovery. Specifically, computational approaches based on cheminformatics and artificial intelligence are being used to facilitate natural product based drug discovery [22, 23]. In this regard, the exploration and characterization of natural products from diverse origins such as bacteria, fungi, marine organisms, plants, etc. will be of immense use. Among the above sources, plants and fungi have contributed to several therapeutic natural products which have been successfully developed as drugs [24]. Also, many plants and fungi are considered medicinal and are widely used in traditional systems of medicine practised world-wide for treatment of several human ailments [25]. These medicinal plants and medicinal fungi with known ethnopharmacological use in traditional systems of medicine are a treasure trove for iden-

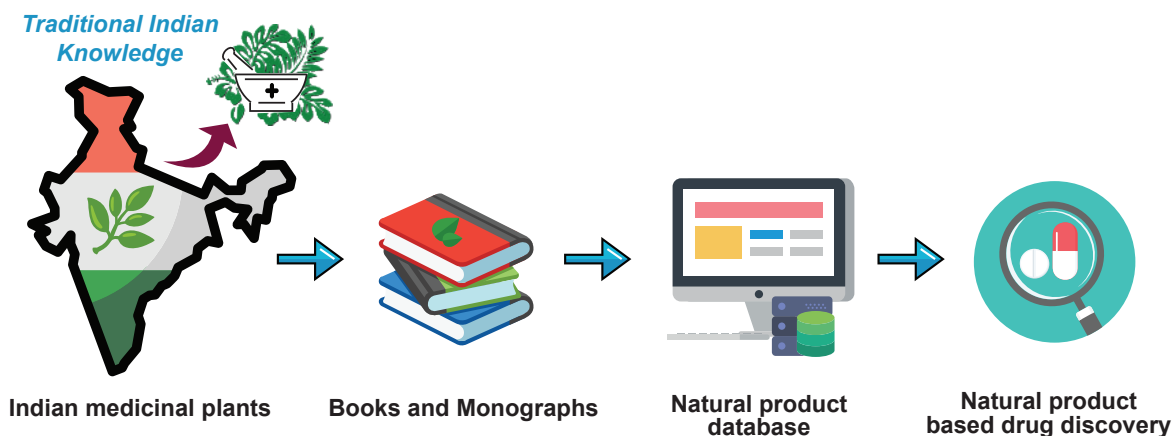


Figure 1.2: Schematic figure shows the potential importance of digitizing traditional knowledge associated with Indian medicinal plants for natural product based drug discovery.

tification of novel therapeutic molecules [26]. Thus, mapping the natural product space of the medicinal plants and medicinal fungi along with their ethnopharmacological uses will enable effective exploration of the biologically relevant and more likely therapeutic chemical space for drug discovery.

India is well known for its practice of traditional medicine and ethnopharmacology [27]. Ayurveda, Siddha and Unani are among the most popular traditional Indian systems of medicine which have gained world-wide recognition for their holistic approach towards treating disease conditions [28]. Apart from them there exists large number of folklore medicine used by traditional communities spread across the length and breadth of India [28]. Many of these traditional systems of medicine specifically Ayurveda, Siddha and Unani make extensive use of medicinal plants indigenous to India in their medicinal formulations or drugs [29]. It is noteworthy that many of the traditional Indian medicinal formulations are multi-component mixtures whose therapeutic use is based on empirical knowledge rather than a mechanistic understanding of the active ingredients in the mixture [27]. Until recently, knowledge of traditional Indian medicine including important medicinal plants and their formulations were buried within books such as Indian Materia Medica [30] and Ayurveda Materia Medica [31]. The non-digital nature of this information limited their effective use towards new drug discovery [21]. Further, molecular mechanisms behind the therapeutic action of medicinal plants used in traditional Indian

medicine remain largely undiscovered. This poses a significant challenge towards turning a largely experience-based enterprise to evidence-based practice, leading to modernization of traditional Indian medicine. Hence, the digitization of this knowledge into a comprehensive database on Indian medicinal plants, phytochemicals and ethnopharmacology will enable researchers to apply computational approaches towards drug discovery (Figure 1.2).

In this direction, there has been significant progress in building databases on natural products with a focus on phytochemicals of edible and herbaceous plants. Examples of such databases include CVDHD [32], KNAPSACK [33], Nutrichem [34,35], Phytochemica [36], SerpentinaDB [37], NPASS [38], CMAUP [39], HIT 2.0 [40], TCM@Taiwan [41], TCMID [42] and TCM-Mesh [43] which can facilitate computational screening of prospective drug compounds or aid in the investigation of plant-disease associations. Yet, from the perspective of traditional Indian medicine, there have been comparatively few efforts to build online databases that capture Indian medicinal plants, their phytochemicals and therapeutic uses. Previously, Polur *et al.* [44] compiled information on 295 ayurvedic Indian medicinal plants, their 1829 phytochemicals and therapeutic uses. Subsequently, Polur *et al.* [44] investigated the chemical structure similarity between their library of 1829 phytochemicals and drugs in the DrugBank [45] database to predict pharmacologically active phytochemicals. The Phytochemica [36] database captures data on 5 Indian medicinal plants and their 963 phytochemicals and the SerpentinaDB [37] database provides information on 147 phytochemicals of a Indian medicinal plant. Other efforts to build online databases for traditional Indian medicine have largely been limited to cataloguing medicinal plants and their therapeutic uses rather than capturing the phytochemicals that are vital for drug discovery. On the other hand, in contrast to the above mentioned online databases, more comprehensive databases are available for Chinese medicinal plants [41–43, 46, 47]. For example, TCM-MeSH [43] is an online database for traditional Chinese medicine which captures phytochemical compositions and therapeutic uses for more than 6000 Chinese medicinal plants. Similarly, there exists databases such

as NPATLAS [48,49] which compile the natural products from bacteria and fungi. Yet, there is no dedicated database capturing the secondary metabolites and therapeutic uses of medicinal fungi (or mushrooms).

In this thesis, we thus first focus on compilation, curation and exploration of the natural product spaces of Indian medicinal plants to enable traditional knowledge based drug discovery. Specifically, we have built a comprehensive manually curated database, Indian Medicinal Plants, Phytochemistry And Therapeutics (IMPPAT) through literature mining followed by manual curation of information gathered from specialized books on traditional Indian medicine, published research articles and other existing online database resources [50,51]. This is followed by construction of the first dedicated database on medicinal fungi, Medicinal Fungi Secondary metabolites And Therapeutics (MeFSAT) [52]. Further, using several cheminformatics and network based approaches we characterize the curated natural product spaces for easier exploration and identification of novel therapeutic molecules. The thesis lastly highlights a specific biological application of the curated phytochemical space of Indian medicinal plants to identify potential anti-COVID lead molecules using computer-aided drug discovery approaches [53,54].

In the subsequent sections of this chapter we provide an overview of the natural product spaces captured in this thesis and description of the various analysis used to characterize them.

1.2 Phytochemical space of Indian medicinal plants

The Indian medicinal plants are a rich source of novel phytochemicals which can enrich and expand the natural product space. Creation of a comprehensive database on Indian medicinal plants, their phytochemicals, their therapeutic uses and their traditional medicinal formulations will be of immense use in natural product and traditional knowledge based drug discovery. We therefore have built a manually curated database, IMPPAT (versions 1.0 and 2.0) [50,51] on Indian medicinal plants, their phytochemicals and their

therapeutic uses. IMPPAT 1.0 provided information on 1742 Indian medicinal plants, 9596 phytochemicals, and 1124 therapeutic uses [50]. In addition, the IMPPAT 1.0 linked Indian medicinal plants to 974 openly accessible traditional Indian medicinal formulations. Importantly, our curation efforts have led to a non-redundant chemical library of 9596 phytochemicals with two-dimensional (2D) and three-dimensional (3D) chemical structures. The phytochemicals in IMPPAT 1.0 were annotated with physicochemical properties, drug-likeness properties and predicted Absorption, distribution, metabolism, excretion and toxicity (ADMET) properties. We also provided predicted interactions between phytochemicals in IMPPAT 1.0 and human target proteins from STITCH [55] database. IMPPAT 1.0 served as the largest resource on Indian medicinal plants and their phytochemicals [9, 50]. Notably, the phytochemical library of IMPPAT 1.0 was used in several computational studies to identify potential lead molecules against variety of diseases including COVID-19 [53, 54, 56–59].

We have subsequently updated IMPPAT 1.0 and have built IMPPAT 2.0, an enhanced and expanded phytochemical atlas of Indian medicinal plants [51]. IMPPAT 2.0 is built upon the published data of earlier version 1.0, and now provides information on 4010 Indian medicinal plants, 17967 phytochemicals, 1095 therapeutic uses and 1133 traditional Indian medicinal formulations [51]. Firstly, 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. Secondly, IMPPAT 2.0 now provides the phytochemical composition, therapeutic uses, and traditional medicinal formulations of Indian medicinal plants at the level of plant parts such as stem, root or leaves. Thirdly, IMPPAT 2.0 provides a FAIR [60] compliant non-redundant *in silico* stereo-aware library of 17967 phytochemicals with 2D and 3D chemical structures. Fourthly, we have characterized the molecular complexity and the molecular scaffold based structural diversity of the phytochemical space of IMPPAT 2.0, and thereafter, compared with other chemical libraries. Fifthly, we have also filtered a subset of 1335 drug-like phytochemicals using multiple drug-likeness rules. Fi-

nally, we have compared the phytochemicals in IMPPAT 2.0 with phytochemicals from Chinese medicinal plants. From our cheminformatics analysis, we find that phytochemicals in IMPPAT 2.0 are more likely enriched with specific protein binders rather than promiscuous binders, have scaffold diversity similar to many larger natural product libraries, and share minimum overlap with the phytochemical space of Chinese medicinal plants. These results reveal the unique features of the phytochemical space of Indian medicinal plants captured in IMPPAT 2.0. The data present in IMPPAT 2.0 is accessible at: <https://cb.imsc.res.in/imppat/>.

1.3 Secondary metabolite space of medicinal fungi

Fungi are present in every ecological niche, and thus, face challenges from myriad biotic and abiotic stressors [61, 62]. Investigation of the fungal habitats has shed important insights into the stimulation and the biosynthesis of valuable secondary metabolites produced by fungi [63]. 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 discovery [64] including antibiotics. Notably, the first broad-spectrum antibiotic, penicillin, discovered by Alexander Fleming is a fungal secondary metabolite.

The fungal kingdom encompasses diverse organisms ranging from simple yeasts to mushrooms. Mushrooms are macrofungi with fruiting bodies [65] that have been used as food and/or medicine for centuries across many civilizations [25, 66–70]. During manual curation of the IMPPAT [50, 51] database on phytochemicals of Indian medicinal plants presented in Chapter 2, we realized that there was no dedicated resource on secondary metabolites of medicinal fungi (or mushrooms) to date. This is surprising given medicinal mushrooms [25, 66–69], have been used for centuries in traditional medicine, especially in many Asian countries [71]. Existing microbial natural product databases such as NPAT-LAS [48, 49] 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 dedicated resource on secondary metabolites and therapeutic uses of medicinal fungi is needed to make full use of this chemical space for drug discovery. We have addressed this unmet need by building a natural product database dedicated to secondary metabolites of medicinal fungi, MeFSAT.

MeFSAT compiles information on 184 medicinal fungi, 1830 secondary metabolites and 149 therapeutic uses [52]. MeFSAT provides the 2D and 3D structures of the 1830 secondary metabolites of medicinal fungi compiled in the database. Further, similar to the IMPPAT database [50, 51], the secondary metabolites are annotated with physicochemical properties, drug-likeness properties, predicted ADMET properties and several other features. From the cheminformatic analysis we find that the secondary metabolites captured in MeFSAT are more likely to be enriched with specific protein binders and they are structurally diverse. Also using multiple drug-likeness rules, we have filtered a subset of 228 drug-like secondary metabolites in MeFSAT database. The compiled information in MeFSAT database can be openly accessed at: <https://cb.imsc.res.in/mefsat/>.

1.4 Cheminformatics based analysis of natural product spaces

The field of cheminformatics deals with the application of computational approaches to analyze, characterize and manipulate the chemical structures in virtual chemical libraries [72]. In the last decade the increase in computing power has led to creation of virtual libraries whose size has run into several billions of molecules [73]. Cheminformatics based methods have become quintessential in exploration of such gargantuan chemical spaces. Cheminformatics has also been used specifically for characterization and analysis of the natural product space. In a recent review, Chen and Kirchmair [22] have provided an in-depth discussion on cheminformatics methods used in natural product based drug discovery. We list below some of the important applications of cheminformatics for the

analysis of natural product space for drug discovery.

Firstly, several cheminformatics methods are widely used in the creation of *in silico* natural product libraries. In the process of compilation and curation of the natural product spaces of Indian medicinal plants and medicinal fungi presented in this thesis, we extensively made use of RDKit [74], a open-source cheminformatics python package and Open Babel [75], a open-source chemistry toolbox for chemical structure format conversion, 3D structure generation and structure similarity comparison. Secondly, there exists several chemical databases such as ZINC [76] providing commercial sources for chemicals which can enable physical sourcing of the natural products. To aid in easier purchase of the phytochemicals in IMPPAT, we have linked the phytochemicals to chemical databases such as ZINC [76] and MolPort (<https://www.molport.com/>) using UniChem, [77], a cheminformatics resource which enables cross-referencing between number of chemical databases. Thirdly, cheminformatics tools are used for the computation of physicochemical and other chemical properties of natural product libraries. Specifically, we used RDKit [74] to compute the molecular complexity and physicochemical properties of the natural product spaces captured in this thesis to highlight their unique features. Also, using rule based filters on physicochemical properties, we have identified drug-like natural products. Fourthly, the structural diversity of the natural product libraries can be assessed using concept of molecular scaffold [78]. We have computed molecular scaffold for the phytochemicals in IMPPAT using RDKit [74] and have found its scaffold diversity to be similar to other natural product libraries. Fifthly, the chemical libraries can be computationally classified based on structural features. We have used ClassyFire web-server [79], NP classifier web-server [80] and NP-likeness score [81, 82] to classify the phytochemicals in IMPPAT. Specifically, NP classifier was used to provide natural product specific classification of the phytochemicals and NP-likeness score was used for assessing the natural product likeness of the phytochemicals. Sixthly, using chemical similarity metrics such as Tanimoto coefficient (T_c) [83] computed based on molecular fingerprints, the natural product libraries can be compared with other chemical libraries. We used Tanimoto

coefficient to construct chemical similarity networks of natural product spaces presented in this thesis and also compared the natural products with approved drugs. These analysis also help understand the structural diversity of the studied natural product library and help in the visualization of the natural product space.

Apart from the above listed purposes, cheminformatics methods have been also used for prediction of biological target and bioactivity of the natural products, and de novo designing of natural product inspired molecules [22]. We used computational methods such as molecular docking, protein - ligand interaction prediction and molecular dynamics simulations to identify potential anti-COVID lead molecules from phytochemicals in IMPPAT. The above list of applications of cheminformatics for natural product based drug discovery is by no means exhaustive. Given the distinct chemical features of natural products, several natural product specific cheminformatics tool have been recently developed [22]. Overall, we expect cheminformatics methods and tools will continue to play a key role in better characterization and analysis of novel natural product spaces.

1.5 Thesis organization

The remaining chapters of this thesis are organized as follows.

Chapter 2 presents the compilation and curation of the natural product space of the Indian medicinal plants. IMPPAT is a curated database of Indian Medicinal Plants, Phytochemistry And Therapeutics. IMPPAT version 1.0 (IMPPAT 1.0) provided detailed information on 1742 Indian medicinal plants and 27074 plant-phytochemical and 11514 plant-therapeutic use associations. In particular, IMPPAT 1.0 provided a non-redundant chemical library of 9596 phytochemicals from Indian medicinal plants with standard chemical identifiers and chemical structure information. Subsequently, we built IMPPAT 2.0 an enhanced and expanded phytochemical atlas of Indian medicinal plants, which provides manually curated information on 4010 Indian medicinal plants, 17967 phytochemicals, 1095 therapeutic uses and 1133 traditional Indian medicinal formulations, encom-

passing 189386 plant-part-phytochemical, 89733 plant-part-therapeutic use, and 7815 plant-part-traditional medicinal formulation associations. Notably, IMPPAT 2.0 compiles associations at the level of plant parts, and provides a non-redundant *in silico* stereo-aware library of 17967 phytochemicals from Indian medicinal plants. Altogether, IMPPAT 2.0 is the largest phytochemical atlas of Indian medicinal plants which is accessible without any login or registration requirement at: <https://cb.imsc.res.in/imppat/>. **The work reported in this chapter is contained in the published manuscript [50] and the manuscript [51].**

Chapter 3 presents an in-depth analysis of the enhanced and expanded phytochemical atlas of Indian medicinal plants compiled in IMPPAT 2.0. We characterized the molecular complexity and molecular scaffold based structural diversity of the phytochemical library in IMPPAT 2.0, and have compared it with other small molecule and natural product chemical libraries. From our analysis, we find that phytochemicals in IMPPAT 2.0 have high values of stereochemical complexity and shape complexity similar to a representative natural product library than with commercial or diversity-oriented synthesis libraries provided by Clemons *et al.* [84]. This indicates that phytochemical library of IMPPAT 2.0 is more likely to be enriched with specific protein binders than promiscuous binders. We also find that the phytochemicals in IMPPAT 2.0 are structurally diverse with scaffold diversity similar to many larger natural product libraries. Specifically, we find the scaffold diversity of phytochemicals in IMPPAT 2.0 and other natural product libraries lie in between the scaffold diversity of 100 million organic compounds from PubChem (low diversity) and approved drugs (high diversity). We also filtered based on multiple drug-likeness scores a subset of 1335 drug-like phytochemicals in IMPPAT 2.0. We find that the majority of the drug-like phytochemicals in IMPPAT 2.0 have no similarity to existing approved drugs. Finally, by comparing the phytochemical library of IMPPAT 2.0 with phytochemicals from Chinese medicinal plants, we show that there is minor overlap between the two chemical spaces. In summary, these results highlight the unique features of the phytochemical space of Indian medicinal plants in IMPPAT 2.0.

The work reported in this chapter is contained in the published manuscript [50] and the manuscript [51].

Chapter 4 presents the compilation, curation and exploration of the natural product space of secondary metabolites produced by medicinal fungi. Medicinal Fungi Secondary metabolites And Therapeutics (MeFSAT) is a comprehensive manually curated database which provides information on 184 medicinal fungi, 1830 secondary metabolites and 149 therapeutic uses. Importantly, MeFSAT provides a non-redundant *in silico* natural product library of 1830 secondary metabolites along with information on their chemical structures. We computed the stereochemical complexity and shape complexity of the secondary metabolites in MeFSAT and compared it with the small molecule libraries provided by Clemons *et al.* [84]. We find that the secondary metabolites have high stereochemical complexity and shape complexity similar to a representative natural product library than with commercial or diversity-oriented synthesis libraries. This finding which is similar to the finding for the phytochemical library of IMPPAT highlights that the secondary metabolites of medicinal fungi contained in MeFSAT are also more likely to be enriched with specific protein binders than promiscuous binders. Further, based on multiple drug-likeness scores, we filtered a subset of 228 drug-like secondary metabolites to prioritize the secondary metabolites which can be taken up for drug discovery. We also constructed chemical similarity networks (CSNs), first for all the secondary metabolites and second for the drug-like secondary metabolites in MeFSAT. We find that both the CSNs are very sparse with several disconnected clusters and a large number of isolated nodes. This highlights the structural diversity among the secondary metabolites in MeFSAT. Lastly, by comparing the secondary metabolites with approved drugs, we find that only 82 of 1830 secondary metabolites and 6 of 228 drug-like secondary metabolites in MeFSAT are structurally similar to any of the approved drugs. In summary, the above findings underscore the diversity of the secondary metabolites of medicinal fungi captured in MeFSAT and their potential to further enrich the natural product space. The compiled information in MeFSAT database can be openly accessed at: <https://cb.ims.res.in/mefsat/>.

The work reported in this chapter is contained in the published manuscript [52].

Chapter 5 presents a biological application of the curated natural product space of Indian medicinal plants captured in IMPPAT for the identification of potential anti-COVID drugs. In the first study, we identified potential phytochemical inhibitors of key host factors, Transmembrane Protease Serine 2 (TMPRSS2) and cathepsin L, which play an important role in SARS-CoV-2 host cell entry [85, 86]. In the second study, we identified potential phytochemical inhibitors targeting the ATP binding site of SARS-CoV-2 helicase Nsp13 which is a promising target for developing anti-COVID drugs [87]. We used (a) the binding energy from molecular docking of the phytochemicals to the active site of the target proteins and (b) the ligand binding site residues and non-covalent interactions between protein and ligand to filter and identify phytochemical inhibitors that either bind to or form interactions with residues important for the specificity/activity of the target proteins. Altogether, we identified 96 inhibitors of TMPRSS2 and 9 inhibitors of cathepsin L among phytochemicals of Indian medicinal plants. Similarly, we identified 368 phytochemicals as potential inhibitors of SARS-CoV-2 helicase Nsp13. The potential inhibitors identified in these two studies can be taken up in the drug discovery pipeline for the development of anti-COVID drugs after *in vitro* and *in vivo* experiments. **The work reported in this chapter is contained in the published manuscripts [53, 54].**

Chapter 6 provides a brief summary of the research on multiple natural product spaces and their biological application for drug discovery presented across chapters of this thesis. The chapter also discusses the future directions based on research presented in this thesis.

Chapter 2

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

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 [88, 89]. Phytochemicals of medicinal plants encompass a diverse natural product space for drug discovery. India is rich with a flora of indigenous medicinal plants that have been used for centuries in traditional Indian medicine to treat human maladies. Until recently, knowledge of traditional Indian medicine including important medicinal plants and their formulations were buried within books and monographs. The non-digital nature of this information limits its complete and effective use in drug discovery research. A comprehensive online database on the phytochemicals of Indian medicinal plants will enable computational approaches towards natural product based drug discovery.

Towards this goal, we built the manually curated database, IMPPAT (version 1.0) [50], containing 1742 Indian medicinal plants, their 9596 phytochemicals, and their therapeu-

tic uses. Importantly, IMPPAT 1.0 compiled two dimensional (2D) and three dimensional (3D) chemical structures of the phytochemicals in the database, along with their physico-chemical, drug-likeness, and absorption, distribution, metabolism, excretion and toxicity (ADMET) properties. 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 [53, 54, 56, 57, 90, 91]. Given the widespread use of IMPPAT 1.0, we subsequently built IMPPAT 2.0, an enhanced and expanded phytochemical atlas of Indian medicinal plants. The latest update, IMPPAT 2.0, built upon the published data of earlier version [50], and now compiles information on 4010 Indian medicinal plants, 17967 phytochemicals, 1095 therapeutic uses and 1133 traditional Indian medicinal formulations. IMPPAT 2.0 is accessible without any login or registration requirement via a user friendly web-interface at: <https://cb.imsc.res.in/imppat/>.

In this chapter, we present the workflow for building IMPPAT 1.0, followed by detailed description on the update to create IMPPAT 2.0. **The work reported in this chapter is contained in the published manuscript [50] and the manuscript [51].**

2.1 Workflow for construction of IMPPAT 1.0

2.1.1 Curated list of Indian medicinal plants

In the preliminary phase of the IMPPAT 1.0 database construction (Figure 2.1), we compiled a comprehensive list of more than 5000 Indian medicinal plants based on information contained in the Indian medicinal plants database (<http://www.medicinalplants.in/>) of the Foundation for Revitalisation of Local Health Traditions (FRLHT), Bengaluru. In addition to the comprehensive list from FRLHT, the AYUSH priority list was compiled from two sources, namely, the list prepared by the National Mission on Medicinal Plants of Ministry of AYUSH, Government of India, and the list jointly prepared by the National Medicinal Plants Board (NMPB), Directorate of Medicinal and Aromatic Plants Research (DMAPR), Department of Agriculture and Cooperation, and

Central Institute of Medicinal and Aromatic Plants (CIMAP), Government of India which is available at <http://pib.nic.in/newsite/PrintRelease.aspx?relid=67277>. We remark that the AYUSH priority list was prepared based on several criteria including the medicinal use, conservation status and herbal industry demand of Indian medicinal plants. Due to the usage of multiple synonyms for Indian medicinal plants across different sources, the common names of plants were manually mapped to their scientific species names using The Plant List database [92] (<http://www.theplantlist.org/>), and the compiled list was manually curated to remove redundancies. Furthermore, the Indian medicinal plants in our database were manually classified into their respective taxonomic families within the kingdom plantae using The Plant List database [92] and Tropicos database (<http://www.tropicos.org/>). We have also linked the Indian medicinal plants in IMPPAT database to their corresponding page in The Plant List database, Tropicos database and the FRLHT digital herbarium (<http://envis.frlht.org>).

2.1.2 Phytochemical composition of Indian medicinal plants

After compiling a comprehensive list of more than 5000 Indian medicinal plants, we mined literature to gather information on their phytochemicals (Figure 2.1). In the first stage of data mining, we focussed on specialized traditional Indian medicine books [93–102]. From these books [93–102], we gathered phytochemical composition for more than 1600 Indian medicinal plants. In the second stage, we gathered information from published databases of Indian medicinal plants. Phytochemica [36] database contains information on 963 phytochemicals of 5 Indian medicinal plants. Another database described in Polur *et al.* [44] had compiled information on 1829 phytochemicals of 295 ayurvedic Indian medicinal plants [44]. While this list is no longer publicly available, the Nutrichem [34, 35] database on phytochemical composition and therapeutic uses of plant-based food products has incorporated the information compiled by Polur *et al.* [44]. From the Phytochemica [36] and Nutrichem [34, 35] databases, we gathered information on the phytochemical composition of more than 400 Indian medicinal plants. Note

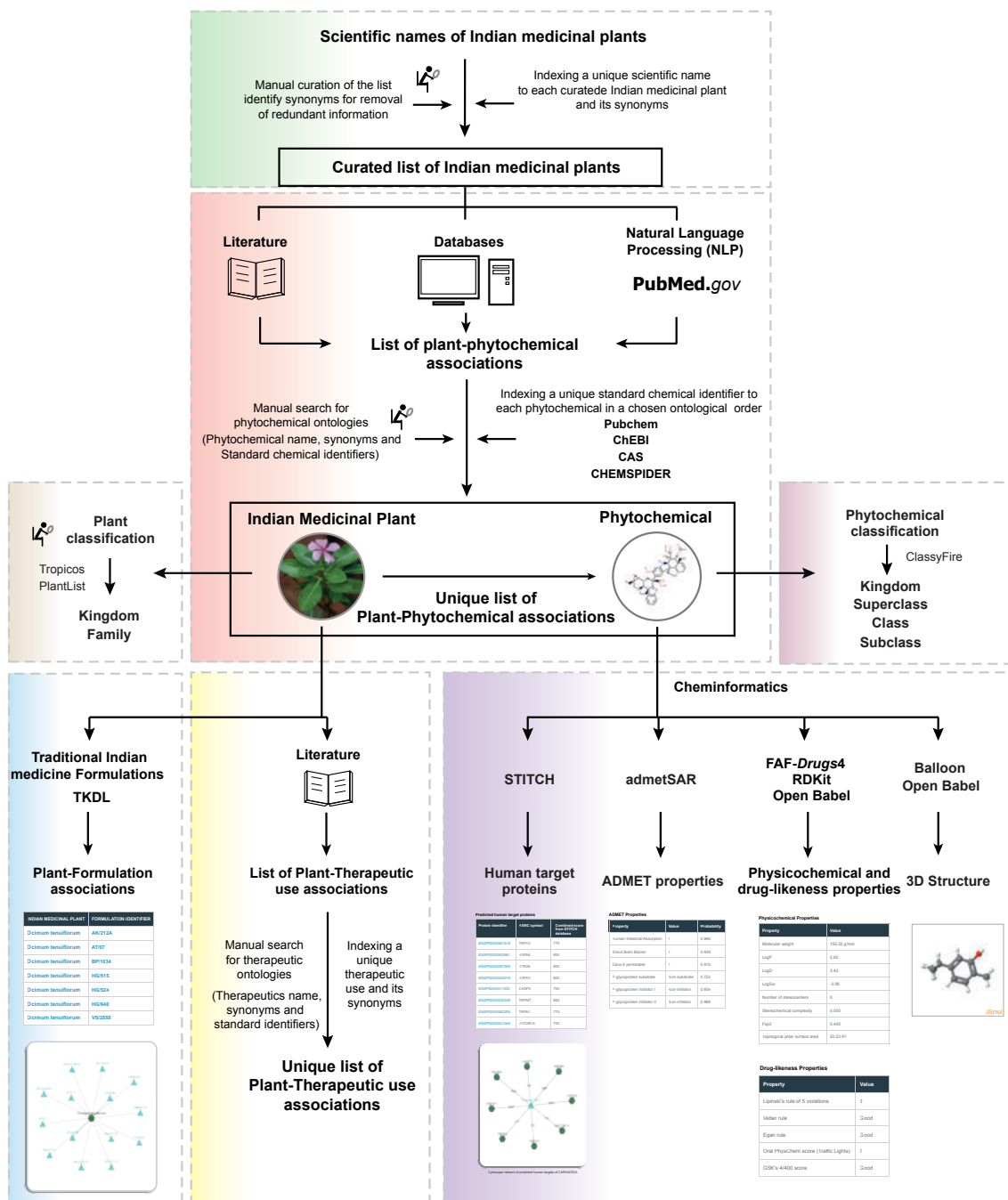


Figure 2.1 (previous page): Schematic overview of the IMPPAT 1.0 database construction pipeline. Briefly, we first compiled a comprehensive list of Indian medicinal plants from various sources. We next mined specialized books on Indian traditional medicine, existing databases and PubMed abstracts of journal articles to gather information on phytochemicals of Indian medicinal plants. We then manually annotated, curated and indexed names of identified phytochemicals with standard identifiers to build a non-redundant library of phytochemicals. This manual curation effort led to a unique list of plant-phytochemical associations. We also classified the Indian medicinal plants into taxonomic families and phytochemicals into chemical classes. Subsequently, we gathered ethnopharmacological information from books on traditional Indian medicine to build a unique list of plant-therapeutic use associations. We also extracted publicly accessible information on traditional medicine formulations from TKDL database to build a list of plant-formulation associations. Lastly, we have used cheminformatic tools to obtain the 3D structures, physicochemical properties, drug-likeness properties, predicted ADMET properties and predicted target human proteins of phytochemicals.

that our comprehensive list covers a wide spectrum of Indian medicinal plants which includes apart from Ayurveda, other systems of traditional Indian medicine such as Siddha and Unani. In the third stage of data mining for phytochemical composition, we performed text mining of abstracts from published research articles in PubMed [103] using natural language processing (NLP) [104]. Using in-house Python scripts and a dataset of known plant-phytochemical associations, we identified keywords in PubMed abstracts which imply plant-phytochemical associations (Supplementary Table S2.1). We then used the selected keywords listed in Supplementary Table S2.1 to mine PubMed abstracts to identify and incorporate additional references for plant-phytochemical associations in our database. In total, our database captures the phytochemical composition of 1742 Indian medicinal plants. The literature references for plant-phytochemical associations are listed in our database in the form of ISBN or DOI identifiers for books and PubMed identifiers (PMIDs) for journal articles.

We would like to mention a potential bias in the list of plant-phytochemical associations compiled from scientific literature. Our database most-likely contains high-quality yet incomplete information on phytochemical composition of Indian medicinal plants. That is, phytochemicals listed are most-likely produced by the corresponding Indian medicinal plant but other phytochemicals not listed in our database cannot be ruled out from being also produced by the same plant due to possible lack of scientific liter-

ature. Moreover, the scientific literature will most probably have more information on phytochemical composition of well-studied or sequenced Indian medicinal plants such as *Catharanthus roseus*. Nevertheless, one can argue that for the discovery of novel molecules it is more important to know the list of phytochemicals produced by an herb rather than the list of phytochemicals not produced by an herb.

2.1.3 Annotation, curation and filtering of identified phytochemicals

An overarching goal of this work is to create a platform for exploring the chemistry of the phytochemicals of Indian medicinal plants. We would like to emphasize that synonymous chemical names are pervasive across the literature on traditional Indian medicine which were mined to construct this database. In order to remove redundancy, we manually annotated the common names of phytochemicals of Indian medicinal plants compiled from literature sources with documented synonyms and standard chemical identifiers (Figure 2.1) from PubChem [105], CHEBI [106], CAS (<https://www.cas.org/>), CHEMSPIDER [107], KNAPSACK [108], CHEMFACES (<http://www.chemfaces.com>), FOODB (<http://foodb.ca/>), NIST Chemistry webbook [109] and Human Metabolome database (HMDB) [110]. While assigning standard identifiers to phytochemicals in our database, we have chosen the following priority order: PubChem [105], CHEBI [106], CAS, CHEMSPIDER [107], KNAPSACK [108], CHEMFACES, FOODB, NIST Chemistry webbook [109] and HMDB [110]. We highlight that this extensive manual curation effort led to the mapping of more than 15000 common names of phytochemicals used across literature sources to a unique set of 9596 standard chemical identifiers. Phytochemicals which could not be mapped to standard chemical identifiers were excluded from our finalized database. Our choice to include only phytochemicals with standard identifiers and structure information was dictated by our goal to investigate the chemistry and drug-likeness of phytochemicals of Indian medicinal plants. We remark that the 2D structure information for the 9596 IMPPAT phytochemicals was obtained using the standard chemical identifiers from the respective databases. We have also deter-

mined the chemical classification of the IMPPAT phytochemicals using ClassyFire [79] (<http://classyfire.wishartlab.com/>). ClassyFire [79] gives a hierarchical classification for each chemical compound into kingdom (organic or inorganic), followed by superclass, followed by class, followed by subclass. Note that ClassyFire classifies organic compounds into 26 superclasses. In a nutshell, this largely manual effort led to compilation of a non-redundant chemical library of 9596 phytochemicals of Indian medicinal plants with standard identifiers and structure information.

2.1.4 Therapeutic uses of Indian medicinal plants

Another goal of our database is to compile ethnopharmacological information on Indian medicinal plants. Towards this goal, we manually compiled the medicinal (therapeutic) uses of Indian medicinal plants from books on Indian traditional medicine [93–102, 109, 111–124]. Apart from books, Polur *et al.* [44] had previously compiled a list of therapeutic uses for 295 ayurvedic Indian medicinal plants, and this information was extracted from the Nutrichem [34, 35] database. To ensure high quality, we manually curated information on therapeutic uses of Indian medicinal plants and consciously avoided automated text mining to retrieve additional information on plant-therapeutic use associations. We remark that our database has manually compiled therapeutic uses of Indian medicinal plants from standard books on traditional Indian medicine which contain accumulated experience-based knowledge on treating human diseases. Furthermore, we manually annotated and standardized the compiled therapeutic uses of Indian medicinal plants from the above sources with standard terms and identifiers from the Disease Ontology [125], Online Mendelian Inheritance in Man (OMIM) [126], Unified Medical Language System (UMLS) [127] and Medical Subject Headings (MeSH) [128] databases. To the best of our knowledge, this is the first large-scale attempt to link the ethnopharmacological information on Indian medicinal plants with standardized vocabulary in modern medicine. Note that databases of gene-disease associations [129] and disease-symptom associations [130] usually provide disease information in form of identifiers from UMLS

and MeSH databases, and in future, information from such databases can be effortlessly integrated into the database.

2.1.5 Traditional formulations of Indian medicinal plants

Traditional knowledge digital library (TKDL) (<http://www.tkd1.res.in>) is a knowledgebase of traditional Indian medicinal formulations. A traditional medicinal formulation is often a multi-component mixture derived from plant, animal and other sources which is used for treating disease based on specific indication. For example, Thinavu Sori Soolaiiku Ennai (TKDL Identifier: HM02/36) is a medicinal formulation in traditional Indian system of medicine, Siddha, which is used to treat allergic rashes, and this formulation mainly consists of extracts of medicinal plants, *Plumbago zeylanica*, *Sesamum orientale* (also called *Sesamum indicum*) and *Cuminum cyminum*. At the time of construction of IMPPAT 1.0, TKDL had more than 250000 formulations of Ayurveda, Siddha and Unani of which 1200 representative formulations are openly accessible via their database. To exhibit the broader utility of our database to phytopharmacology, we have also compiled and curated the subset of 1200 openly accessible formulations in TKDL which contain at least one of the 1742 Indian medicinal plants in our database. This process led to associations between 321 Indian medicinal plants in our database and 974 traditional Indian medicinal formulations which are openly accessible through TKDL database (Figure 2.1).

In summary, we constructed IMPPAT 1.0, a curated database of 1742 Indian Medicinal Plants, 9596 phytochemicals, and 1124 therapeutic uses spanning 27074 plant-phytochemical associations and 11514 plant-therapeutic use associations (Table 2.1). IMPPAT 1.0 also provides chemical classification, 2D and 3D chemical structure, and other structure information for the phytochemicals (Table 2.1). Finally, IMPPAT 1.0 captures limited information on the associations between Indian medicinal plants and their use in traditional Indian medicinal formulations. Table 2.1 provides a comparison of the IMPPAT 1.0 database with previous efforts by Polur *et al.* [44] and Phytochemica [36]

to build dedicated digital resource on phytochemical composition of Indian medicinal plants.

2.2 IMPPAT 2.0: an enhanced and expanded phytochemical atlas of Indian medicinal plants

In this section, we present the detailed description on the updated version 2.0 of IMPPAT, which is a significant enhancement and expansion over the previous version 1.0 (Table 2.2). This update was realized through extensive manual curation and addition of several new features to IMPPAT (Figure 2.2; Table 2.2). Figure 2.2 summarizes the important features including enhancements accomplished in IMPPAT 2.0.

2.2.1 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 more than 2-fold increase in the coverage of medicinal plants with respect to the previous version (Table 2.2). During data collection from various sources, we encountered extensive use of synonymous plant names in 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 pharmacopoeia. For this reason, IMPPAT 2.0 provides the compiled information for a non-redundant list of 4010 Indian medicinal plants. This non-redundant 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 published literature in IMPPAT 1.0 or during this update. Second, The Plant List database (<http://www.theplantlist.org/>) was used to identify the accepted scientific names for the compiled plant names. Third, the synonymous names

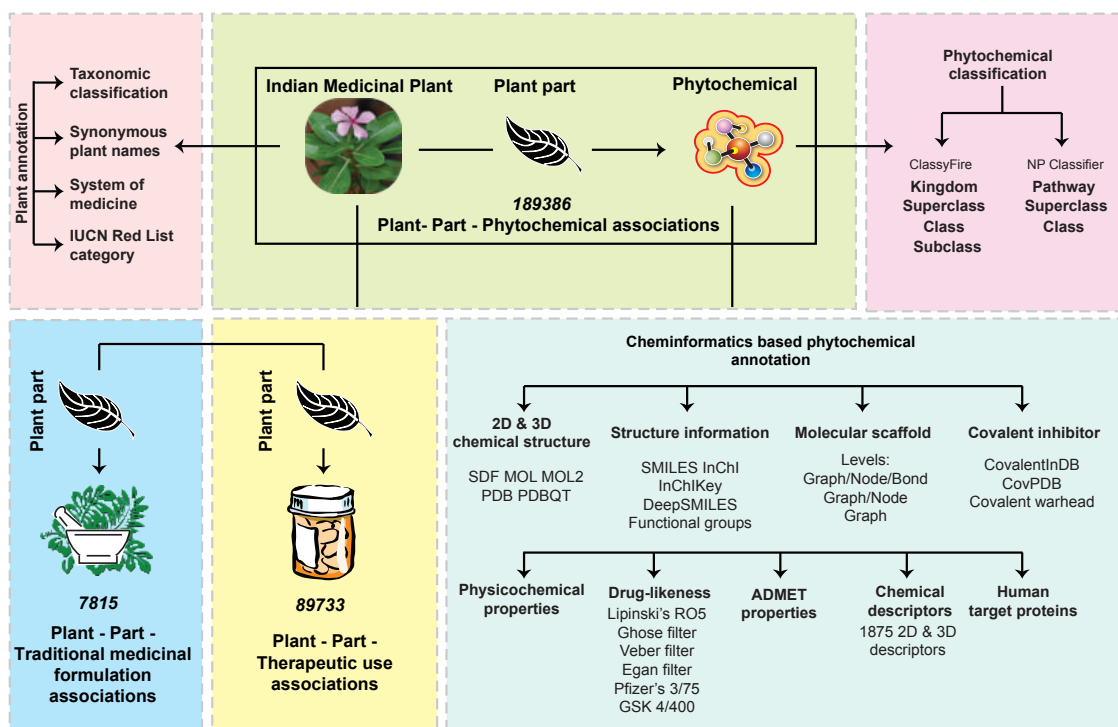


Figure 2.2: Schematic overview of the important features including enhancements and expansion realized in IMPPAT 2.0.

were merged using the accepted scientific names.

Further, the 4010 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 [131]. For the Indian medicinal plants in IMPPAT 2.0, the taxonomic information on kingdom, family and group was compiled using The Plant List database (<http://www.theplantlist.org/>). The common names of the Indian medicinal plants were obtained from the Flowers of India database (<http://www.flowersofindia.net/>), which compiles information for more than 6000 Indian plants. The IUCN Red List of Threatened species [131] (<https://www.iucnredlist.org/>) is the most comprehensive resource on 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 Home-

opathy was manually compiled from pharmacopoeias published by Government of India and Traditional Knowledge Digital Library (TKDL; <http://www.tkdil.res.in>) of the Council of Scientific and Industrial Research, Government of India. The Indian medicinal plants covered in IMPPAT 2.0, have also been provided with cross-reference links to associated information in other standard databases such as The Plant List, Tropicos (<https://www.tropicos.org/>), Encyclopedia of Indian medicinal plants from FRLHT (<http://envis.frlht.org/>), Medicinal Plants Names Service (MPNS; <https://mpns.science.kew.org/>), International Plant Names Index (IPNI; <https://www.ipni.org/>), Plants of the World Online (POW; <https://powo.science.kew.org/>), World Flora Online (WFO; <http://www.worldfloraonline.org/>) and Gardeners' World (<https://www.gardenersworld.com/>).

The 4010 Indian medicinal plants in IMPPAT 2.0 belong to 244 taxonomic families, and Figure 2.3A 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 [132]. 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 2.3B). 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 2.3C). Precariously, we find many of the Indian medicinal plants require extensive conservation effort as 72, 50, 40, 11 and 3 plants are categorized in the IUCN Red list of threatened species as vulnerable (VU), near threatened (NT), endangered (EN), critically endangered (CR), and extinct in the wild (EW), respectively (Figure 2.3D).

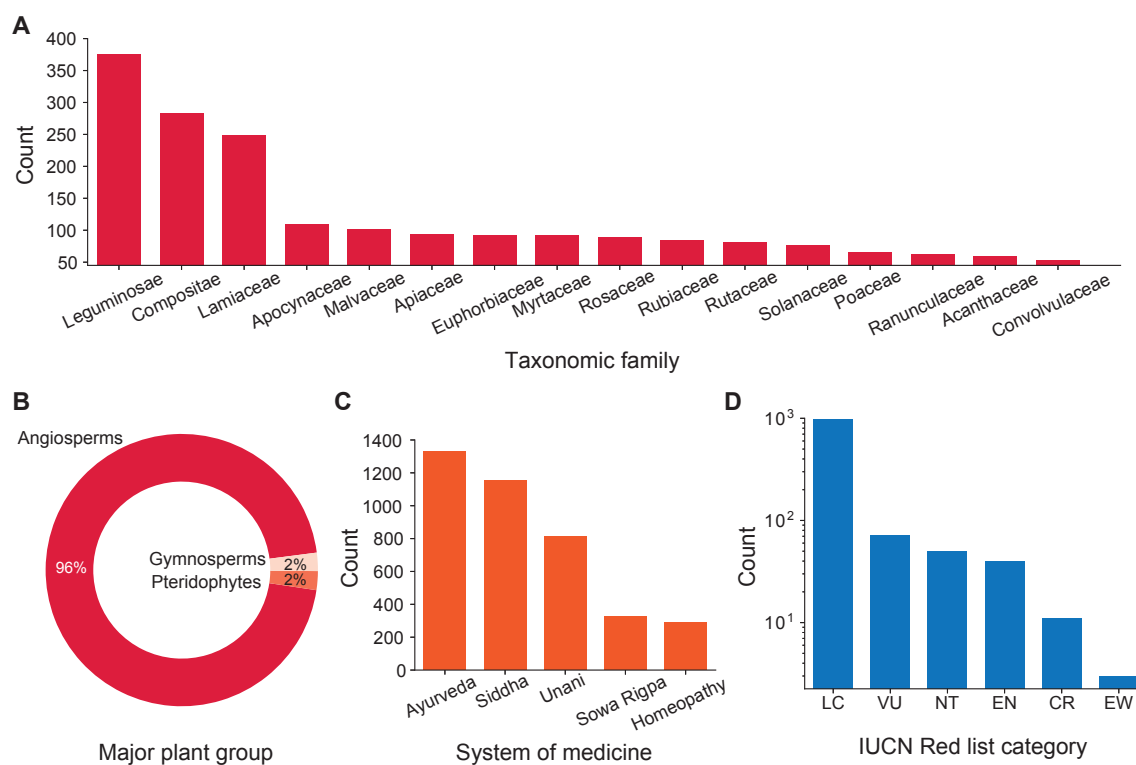


Figure 2.3: Coverage of Indian medicinal plants in IMPPAT 2.0. (A) Top taxonomic families of Indian medicinal plants in IMPPAT 2.0. Note only families with more than 50 Indian medicinal plants in the database are shown. (B) Classification of the Indian medicinal plants in IMPPAT 2.0 into major plant groups: Angiosperms (Flowering plants), Gymnosperms (Conifers, cycads and allies) and Pteridophytes (Ferns and fern allies). (C) Use of Indian medicinal plants in IMPPAT 2.0 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 according to conservation status of the Indian medicinal plants in IMPPAT 2.0. LC – Least concern, VU – Vulnerable, NT – Near threatened, EN – Endangered, CR – Critically endangered, EW – Extinct in the wild.

2.2.2 Information at the level of plant parts

Unlike previous version 1.0, IMPPAT 2.0 provides information on plant – phytochemical, plant – therapeutic use, and plant – traditional medicinal formulation associations at the level of plant parts (Table 2.2). 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 stem, root or leaves. Since it is common knowledge that phytochemical composition can significantly vary across different plant parts, this enhancement in IMPPAT 2.0 will facilitate researchers in phytochemistry and pharmacognosy 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 for preparation of medicinal formulations used to treat various diseases. This further underscores the importance of compiled information in IMPPAT 2.0 on therapeutic use and traditional medicinal formulation at the level of plant parts.

2.2.3 Increase in coverage of phytochemicals

Among the major enhancements in IMPPAT 2.0 is the creation of a non-redundant stereo-aware natural product library of 17967 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 1.0 (Table 2.2).

Building upon the published methodology and extensive data compiled in IMPPAT 1.0 [50], 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 (Supplementary Table S2.2). Only 9 out of these 70 books were covered in IMPPAT 1.0. Importantly, the remaining 61 books covered in IMPPAT 2.0 include: (a) 5 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 pharmacopoeias 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 [133]. Second, aside from the books, all the plant – phytochemical associations compiled from various sources in the previous version IMPPAT 1.0 [50] 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 [37] providing phytochemical information for the Indian medicinal plant *Rauvolfia serpentina*.

A major challenge during compilation, curation and digitization of the phytochemical composition of Indian medicinal plants is the large-scale use of non-standard and synonymous names for phytochemicals in books and research articles. Therefore, to create a non-redundant 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 [134] 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 stereochemistry of the phytochemicals using the InChI. These steps led to the creation of a non-redundant stereo-aware chemical library of 17967 phytochemicals which are produced by 4010 Indian medicinal plants with therapeutic uses. Thus, the phytochemical atlas will aid ongoing efforts towards the identification of novel bioactive and therapeutic molecules.

Overall, there are 189386 non-redundant plant – part – phytochemical associations in IMPPAT 2.0 spanning 4010 Indian medicinal plants and 17967 phytochemicals. At the level of plant – phytochemical associations (after ignoring the plant parts), there is a 5-fold increase in IMPPAT 2.0 (Table 2.2). Figure 2.4A shows the occurrence of phytochemi-

cals 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, while 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 ten plants in terms of the compiled information on the number of phytochemicals produced by them.

2.2.4 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 [60], we assign unique IMPPAT identifiers to phytochemicals in the database, and thereafter, the identifiers are annotated with chemical names, chemical classification, structural features, and external links to standard chemical databases (Figure 2.2). Moreover, we provide the chemical structures of phytochemicals in five different file formats. OpenBabel [75] was used to convert the 2D chemical structures of phytochemicals to SDF, MOL and MOL2 file formats. Whereas the images of the 2D structures of phytochemicals were generated using RDKit [74]. The 3D chemical structures of phytochemicals were retrieved from PubChem [134]. 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 ETKDG method and thereafter energy minimizing the structure using MMFF94 force field [74]. The 3D structures of phytochemicals were then converted to SDF, MOL, MOL2, PDB and PDBQT file formats using OpenBabel [75]. Note that IMPPAT 2.0 provides 3D struc-

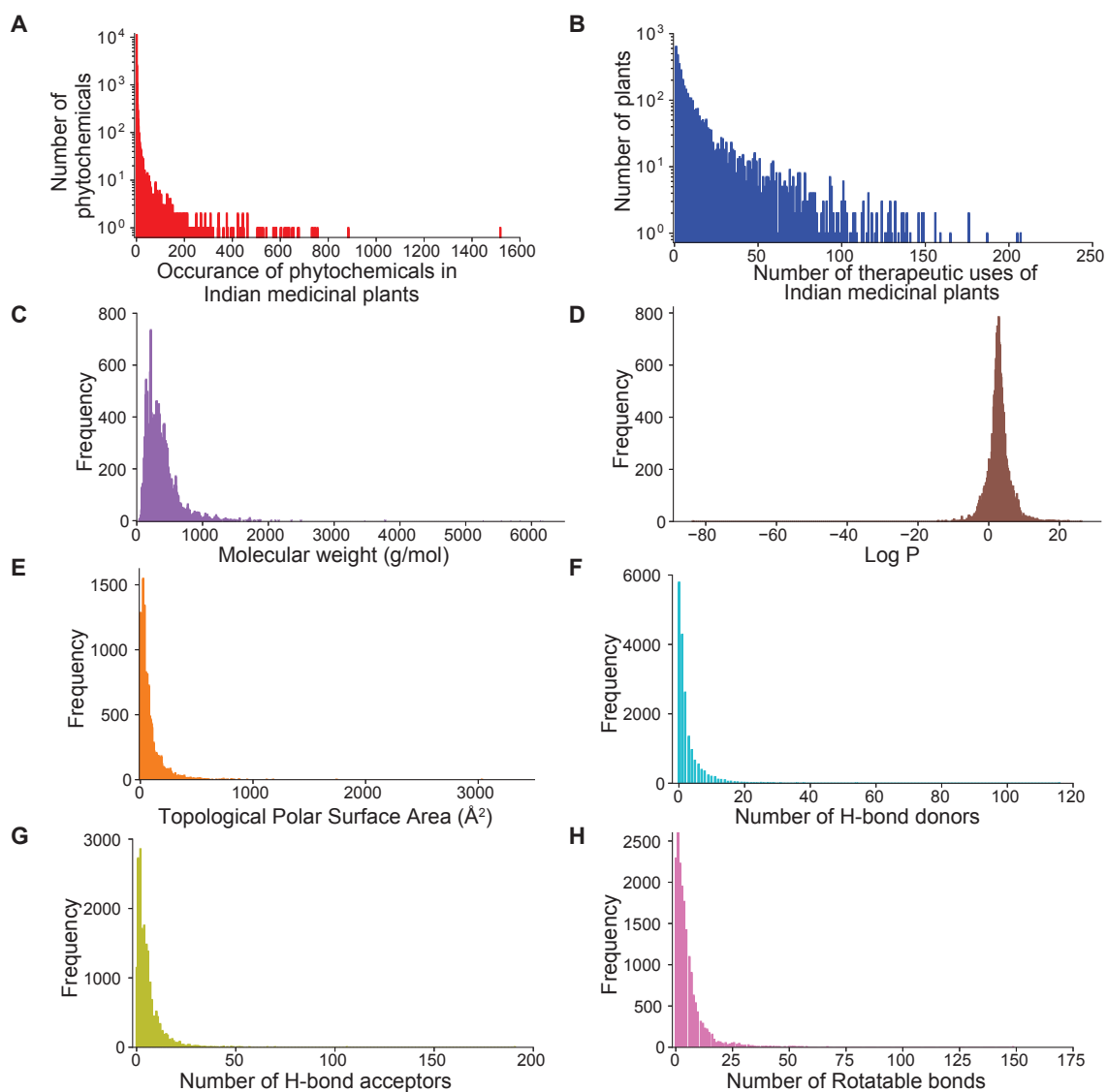


Figure 2.4: Basic statistics and distribution of the physicochemical properties for phytochemicals in IMPPAT 2.0. (A) Histogram of the number of Indian medicinal plants that produce a given phytochemical in IMPPAT 2.0. (B) Histogram of the number of therapeutic uses per Indian medicinal plant in IMPPAT 2.0. Distribution of six important physicochemical properties for 17967 phytochemicals, namely, (C) Molecular weight (g/mol), (D) $\log P$, (E) Topological polar surface area (\AA^2), (F) number of hydrogen bond (H-bond) donors, (G) number of hydrogen bond (H-bond) acceptors, and (H) number of rotatable bonds.

tures for 17910 phytochemicals as the generation of 3D structures failed for the remaining 57 phytochemicals in the database. Lastly, chemical structure of each phytochemical in SMILES, InChI and InChIKey formats was also generated using OpenBabel [75].

Figure 2.4C-H shows the distribution of six important physicochemical properties for the phytochemicals in IMPPAT 2.0. For each phytochemical in our database, the physicochemical properties and drug-likeness scores were computed using in-house custom RDKit scripts. Further, the ADMET properties of the phytochemicals were predicted using SwissADME (<http://www.swissadme.ch/>) [135]. Since the SwissADME restricts the input molecules based on their length of SMILES, therefore, ADMET predictions could not be obtained for 493 phytochemicals in our database.

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.* [136, 137] to compute and provide the molecular scaffolds for phytochemicals at three levels. A detailed description on these scaffolds and their comparison with other chemical libraries is presented in Chapter 3. 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 Peter Ertl [138] to provide the functional groups present in phytochemicals. This functional group information can also facilitate the exploration of the phytochemical space by chemists. Further enhancement of phytochemical annotation in IMPPAT 2.0 include new information such as DeepSMILES [139] which is an adaptation of SMILES for use in machine learning and external links to other standard chemical databases obtained using UniChem [77].

Further, using ClassyFire (<http://classyfire.wishartlab.com/>) [79], the chemical classification for each phytochemical into hierarchical levels namely, kingdom, superclass, class and subclass, was predicted. Based on chemical classification obtained by ClassyFire [79], the 17967 phytochemicals have been hierarchically categorized into 20 superclass, 250 class and 410 subclass. Among the 20 superclass, Lipids and lipid-like molecules, Phenylpropanoids and polyketides, and Organoheterocyclic compounds

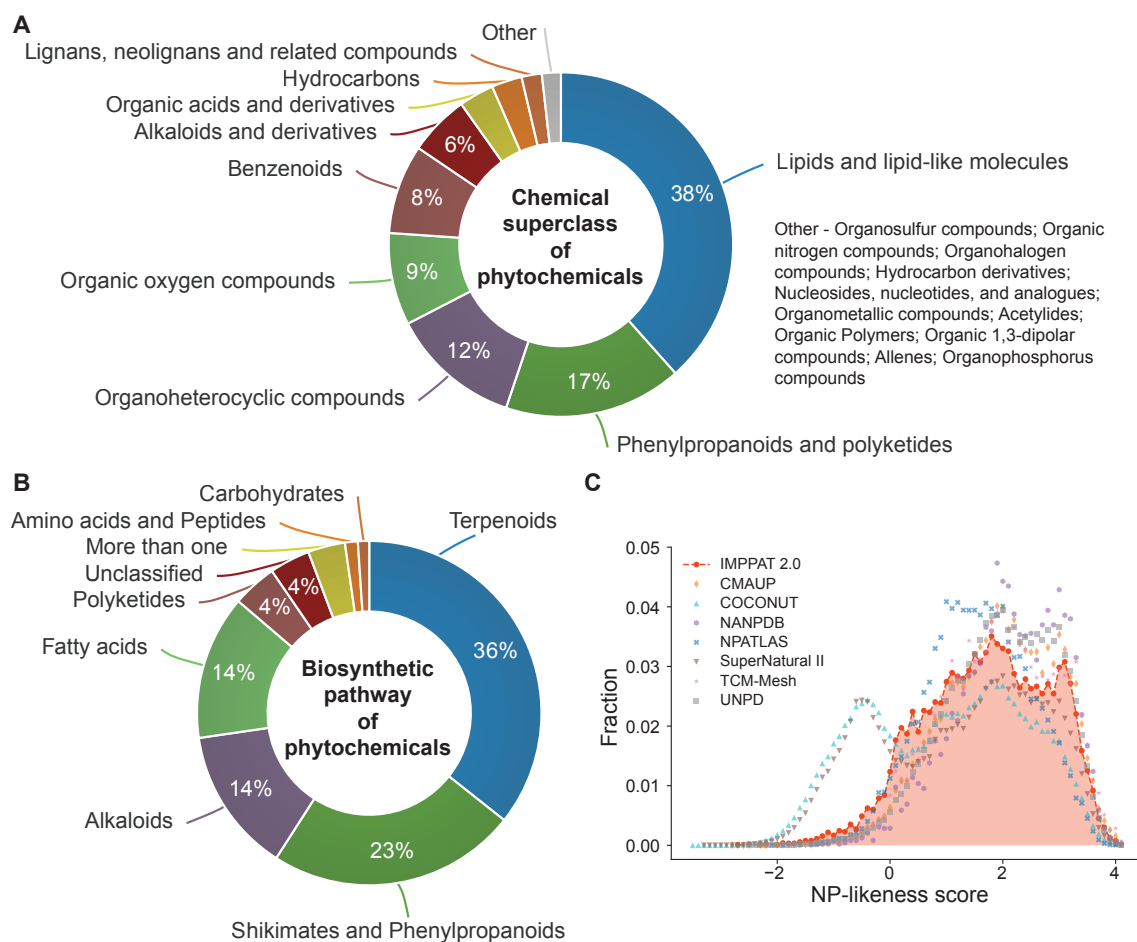


Figure 2.5: Chemical classification, biosynthetic pathways and natural product likeness of phytochemicals in IMPPAT 2.0. (A) Chemical superclass of phytochemicals predicted by ClassyFire. (B) Biosynthetic pathways for phytochemicals predicted by NP classifier. (C) Distribution of the NP-likeness scores for phytochemicals in IMPPAT 2.0 and other natural product libraries.

are the top three with 6904, 3007, and 2202 phytochemicals, respectively (Figure 2.5A). Also using NP classifier (<https://npclassifier.ucsd.edu/>) [80], a natural product specific chemical classification for each phytochemical into biosynthetic pathway, superclass and class was predicted. Based on chemical classification obtained by NP classifier [80], the 17967 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 2.5B).

NP-likeness [82] score is a measure to quantify the similarity of a given chemical structure to the natural product space. This score ranges from -5 to 5; the higher the

score, more likely the molecule is a natural product [82]. Previous studies have shown that the NP-likeness of natural product libraries is predominantly positive, and moreover, is different from synthetic libraries which is predominantly negative [140,141]. We compute the NP-Likeness score for each phytochemical using a custom RDKit script [81,82]. 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 2.5C).

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 using PaDEL [142] software. Lastly, IMPPAT 2.0 compiles information on 27365 predicted interactions between phytochemicals and human target proteins from STITCH [55] database. 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 [143]. These 27365 interactions involve 1294 phytochemicals and 5042 human target proteins

2.2.5 Increase in coverage of therapeutic uses

Building upon the compiled information in IMPPAT 1.0 [50], we enhanced the therapeutic use information in IMPPAT 2.0 to the level of plant parts and expanded 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 (Supplementary Table S2.3). 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 (Supplementary Tables S2.2-S2.3).

Since 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; <https://meshb.nlm.nih.gov/>), International Classification of Diseases 11th Revision (ICD-11; <https://icd.who.int/browse11/>), Unified Medical Language System (UMLS; <https://uts.nlm.nih.gov/uts/umls>) and Disease Ontology (<https://disease-ontology.org/>). 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 non-redundant list of 1095 standardized therapeutic use terms in IMPPAT 2.0.

Overall, there are 89733 unique 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 2.2). Figure 2.4B shows the histogram of the number of therapeutic uses per Indian medicinal plant in IMPPAT 2.0. While 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.

2.2.6 Increase in coverage of traditional medicinal formulations

Finally, IMPPAT 2.0 also compiles information on 7815 plant – part – traditional medicinal formulation associations which encompass 569 Indian medicinal plants and 1133 traditional Indian medicinal formulations (Table 2.2). This information was compiled using 1250 openly accessible formulations in Traditional Knowledge Digital Library (TKDL; <http://www.tkd1.res.in>) database. Further, the 1133 traditional Indian medicinal formulations in IMPPAT 2.0 belong to four systems of medicine, namely, Ayurveda (470), Unani (441), Siddha (187), and Sowa-Rigpa (35).

2.3 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, IMPPAT 1.0 webserver enabled cheminformatics analysis such as filtering phytochemicals based on their physicochemical properties, drug-likeness scores and chemical similarity (Figure 2.6). For the latest release, IMPPAT 2.0, we have completely redesigned the associated website: <https://cb.imsc.res.in/imppat>. 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 in three different ways: (a) Phytochemical association, (b) Therapeutic use, and (c) Traditional medicinal formulation.

The phytochemical association section within browse enables a user to choose either an Indian medicinal plant, or 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-specific information along with a table listing the phytochemical constituents for the plant at the level of plant parts (Figure 2.7A). 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


A Ocimum tenuiflorum

Kingdom: Plantae
Family: Lamiaceae
More Information:

PlantList Tropicos PRLHT

INDIAN MEDICINAL PLANT	PHYTOCHEMICAL IDENTIFIER	PHYTOCHEMICAL NAME	REFERENCES
Ocimum tenuiflorum	CID:12303662	Phytosterols	ISBN:9788171360536
Ocimum tenuiflorum	CID:440966	(-)-camphene	ISBN:9788171360536
Ocimum tenuiflorum	CID:443158	(-)-Linalool	DOI:10.15482/USDA.ADC/1239279
Ocimum tenuiflorum	CID:441005	(+)-delta-Cadinene	DOI:10.15482/USDA.ADC/1239279
Ocimum tenuiflorum	CID:442460	(1S,2R,4S)-(-)-Bornyl acetate	ISBN:9788171360536
Ocimum tenuiflorum	CID:10050	(1S)-1,7,7-Trimethylbicyclo[2.2.1]heptan-2-one	ISBN:9788171360536
Ocimum tenuiflorum	CID:5315468	(E)-alpha-bisabolene	DOI:10.15482/USDA.ADC/1239279
Ocimum tenuiflorum	CID:6424189	1-(4-Hydroxy-3-Methoxyphenyl)-1,2,3-Tris(4-Allyl-2-Methoxyphenoxy)Propane	DOI:10.15482/USDA.ADC/1239279

B CARVACROL



View 3D structure using JSmol

Download structure:

2D: 2D MOL 2D MOL2 2D SDF

3D: 3D MOL 3D MOL2 3D SDF 3D PDB 3D PDBQT

Chemical kingdom: Organic compounds
Superclass: Lipids and lipid-like molecules
Class: Prenol lipids
Subclass: Monoterpenoids
PubChem Identifier: 10364
ChEBI Identifier: 3440
CAS Identifier: 499-75-2
Synonyms: CARVACROL, ACETYLGITOXINS
Canonical SMILES: CC1=C(C)C(C1)O/C/C
InChI Key: RECUKUPTGUEGMW-UHFFFAOYSA-N

C

INDIAN MEDICINAL PLANT	THERAPEUTIC USE	THERAPEUTIC IDENTIFIER	REFERENCES
Ocimum tenuiflorum	ANTIBACTERIAL		ISBN:978-0-387-70637-5
Ocimum tenuiflorum	ANTIFUNGAL	DOI:1564	ISBN:978-0-387-70637-5
Ocimum tenuiflorum	ANTI-PERIODIC		ISBN:978-0-387-70637-5
Ocimum tenuiflorum	ANTI-PYRETIC	DOI:11100	ISBN:978-0-387-70637-5
Ocimum tenuiflorum	ANTISPASMODIC		ISBN:978-0-387-70637-5
Ocimum tenuiflorum	ASTHMA	MESH:D001249, MESH:D001250, MESH:D016538, UMLS:C0004096, UMLS:C0004099, UMLS:C0085129, UMLS:C0155883, DOI:2841	ISBN:978-0-387-70637-5, ISBN:9788171360536
Ocimum tenuiflorum	BRONCHITIS	MESH:D001991, MESH:D029481, UMLS:C0006277, UMLS:C0008677, UMLS:C0148514, UMLS:C2939171, DOI:6132	ISBN:9788171360536

D

Physicochemical properties Drug-like filters Chemical similarity filter

Molecular weight (g/mol) Is equal to []

LogP Is equal to []

Topological polar surface area (Å²) Is equal to []

Number of hydrogen bond acceptor Is equal to []

Number of hydrogen bond donors Is equal to []

Number of heavy atoms Is equal to []

Number of heteroatoms Is equal to []

Number of rigid bonds Is equal to []

Number of rotatable bonds Is equal to []

Stereochemical complexity Is equal to []

Fsp3 Is equal to []

Apply

Select any filter and click on Apply to see results

Physicochemical properties Drug-like filters Chemical similarity filter

Lipinski RO5 violation Oral PhysChem score GSK 4/400 Pfizer 3/75 Veber rule Egan rule QEDw

- Any - - Any - - Any - - Any - - Any - - Any - Is equal to []

Apply

Select any filter and click on Apply to see results

Physicochemical properties Drug-like filters Chemical similarity filter

Enter SMILES *

[]

Choose Fingerprint*

- Select - []

Submit

Figure 2.6 (previous page): Web-interface of the IMPPAT 1.0 database. (A) Snapshot of the result of a standard query for phytochemicals of an Indian medicinal plant. In this example, we show the plant-phytochemical association for *Ocimum tenuiflorum*, commonly known as Tulsi, from IMPPAT 1.0 database. (B) Snapshot of the dedicated page containing detailed information on 2D and 3D chemical structure, physicochemical properties, drug-likeness scores, predicted ADMET properties and predicted target human proteins for a chosen phytochemical. From the dedicated page for each phytochemical, users can download the chemical structure of the phytochemical in the form of a SDF or MOL or MOL2 or PDB or PDBQT file. (C) Snapshot of the result of a standard query for therapeutic uses of an Indian medicinal plant. In this example, we show the therapeutic uses of *Ocimum tenuiflorum* from IMPPAT 1.0 database. (D) Snapshot of the advanced search options which enable users to filter phytochemicals based on their physiochemical properties or drug-likeness scores or chemical similarity with a query compound.

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 2.7B). The users can also retrieve compiled information in IMPPAT 2.0 on the plant – part – traditional medicinal formulation associations by choosing either an Indian medicinal plant, or a TKDL traditional medicinal formulation identifier, or a traditional Indian system of medicine such as Ayurveda, Siddha, Sowa-Rigpa, and Unani.

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 three tabs: (a) Phytochemical association, (b) Therapeutic use, and (c) Traditional medicinal formulation.

In the phytochemical association tab, a user can perform text-based search using complete or partial name of the plant, or IMPPAT phytochemical identifier, or complete or partial name of the phytochemical, to retrieve compiled information in IMPPAT 2.0 on plant – part – phytochemical associations. Upon submitting the text query, the user is presented with a table on the same page listing the relevant plant – part – phytochemical

A

Piper betle

Kingdom: Plantae
Family: Piperaceae
Group: Angiosperms
Common name: Betel Vine
Synonymous names: Piper betel, Piper betle
System of medicine: Ayurveda, Siddha, Sowa Rigpa, Unani

More Information:

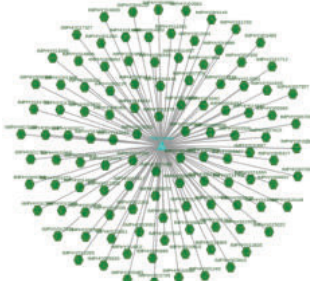
[Plant List](#)
[Tropicos](#)
[World Flora Online](#)
[MPNS Kew](#)
[IPNI](#)
[PoW](#)
[FRLHT Plant Information](#)
[FRLHT Herbarium](#)

Plant - Part - Phytochemical association table

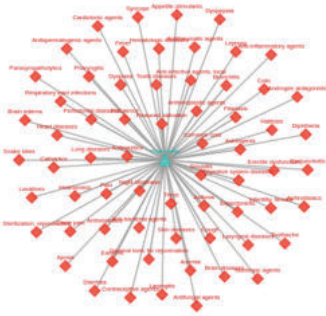
Indian medicinal plant	Plant part	IMPPAT Phytochemical identifier	Phytochemical name	References
Piper betle	leaf	IMPHY017327	p-Menthane-1,3-diol	DOI:10.1002/fj.2730090611

Plant - Part - Therapeutic use association table

Indian medicinal plant	Plant part	Therapeutic use	Therapeutic use identifiers	References
Piper betle	fruit	Cough	MESH:D003371, UMLS:C0010200, ICD-11:MD12	ISBN-9780387706375, ISBN-9788172361266



Plant - Phytochemical association

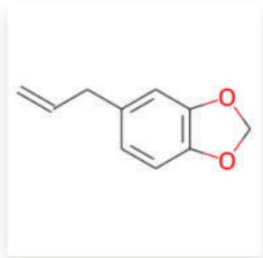


Plant - Therapeutic use association

B

IMPPAT Phytochemical information: Safrole

[Summary](#)
[Physicochemical](#)
[Drug-likeness](#)
[ADMET](#)
[Descriptors](#)
[Predicted human target proteins](#)



[View 3D structure using JSmol](#)

Summary

IMPPAT Phytochemical Identifier: IMPHY004549

Phytochemical name: Safrole

Synonymous chemical names:
 1-allyl-3,4-methylenedioxybenzene, 4-allyl-1,2-(methylenedioxy)-benzene, safrol, safrole

External chemical identifiers:
[CID:5144](#), [ChEMBL:CHEMBL242273](#), [ChEBI:8994](#), [ZINC:ZINC000000002050](#),
[FDASRS:RSB34337V9](#), [SureChEMBL:SCHEMBL56828](#), [MolPort-001-788-008](#)

Figure 2.7: 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 for *Piper betle* 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. (B) Screenshot of the dedicated page containing detailed information for the phytochemical Safrole.

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 complete or partial name of the plant, or therapeutic use term, to retrieve compiled information in IMPPAT 2.0 on plant – part – therapeutic use associations with literature references as a table on the same page. Likewise, in the traditional medicinal formulation tab, a user can perform text-based search using complete or partial name of the plant, or TKDL formulation identifier, to retrieve compiled information in IMPPAT 2.0 on plant – part – traditional medicinal formulation associations as a table on the same page. In this table, on clicking the TKDL formulation identifier, the user is redirected to the corresponding formulation page in TKDL with additional information on the medicinal formulation.

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.

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 which are structurally similar

based on Tanimoto coefficient (T_c) [83], a standard measure to quantify the extent of chemical similarity. 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 scaffold namely, graph/node/bond (G/N/B) level, or graph/node (G/N) level, or graph level (Chapter 3), 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 towards 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 or basic search or advanced search options. The dedicated page provides detailed information for a phytochemical in six tabs: (a) summary, (b) physicochemical, (c) drug-likeness, (d) ADMET, (e) descriptors, and (f) predicted human target proteins (Figure 2.7B). The summary tab provides basic information such as the chemical name, chemical classification, chemical structures, 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 STITCH [55] database, respectively, for the phytochemical. The predicted human target proteins tab also provides a network visualization of the phytochemical – predicted human target protein associations.

2.4 Discussion

IMPPAT is by far the largest phytochemical atlas specific to Indian medicinal plants to date, and this resource is a culmination of our ongoing efforts to digitize the wealth of information contained within traditional Indian medicine. In this chapter we first present the workflow for building IMPPAT version 1.0 (IMPPAT 1.0). IMPPAT 1.0 provided infor-

mation on 1742 Indian medicinal plants, their 9596 phytochemicals and their therapeutic uses. Briefly, for the construction of IMPPAT 1.0, firstly a curated list of Indian medicinal plants was prepared, secondly we compiled information on phytochemicals of the Indian medicinal plants, thirdly we manually curated the phytochemical information to create a non-redundant library of phytochemicals, fourthly we annotated the phytochemicals with features including predicted human target proteins, fifthly we collected and curated information on therapeutic uses of the Indian medicinal plants, and finally we compiled information on traditional medicinal formulations based on use of Indian medicinal plants covered in IMPPAT 1.0.

Subsequently, we present a detailed account of the update to create IMPPAT 2.0. IMPPAT 2.0 is an enhanced and expanded database, compiling information via extensive manual curation on Indian medicinal plants, their phytochemicals, therapeutic uses and traditional medicine formulations. 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, therapeutic uses and traditional medicinal formulations of the Indian medicinal plants at the level of plant parts. At the level of associations, IMPPAT 2.0 compiles 189386 plant – part – phytochemical, 89733 plant – part – therapeutic use, and 7815 plant – part – traditional medicinal formulation associations. Importantly, IMPPAT 2.0 provides a FAIR [60] compliant non-redundant *in silico* stereo-aware library of 17967 phytochemicals. The phytochemical library has been annotated with several features including 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 towards drug discovery. Also, the 1095 standardized therapeutic use terms in IMPPAT 2.0 are mapped to standard terms such as MeSH (<https://meshb.nlm.nih.gov/>), ICD-11 (<https://icd.who.int/browse11/>), UMLS (<https://uts.nlm.nih.gov/uts/umls>) and Disease Ontology (<https://disease-ontology.org/>) used in western medicine. Further, 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. For instance, the advanced search page now enables the user to draw the chemical structure using a visual molecular editor to search for similar phytochemicals in the database and also allows the user to select the phytochemicals based on molecular scaffolds.

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.

Supplementary Information

Supplementary Tables S2.1-S2.3 associated with this chapter is available for download from the GitHub repository: https://github.com/asamallab/PhDThesis-Vivek_Ananth_RP/blob/main/SI/ST_Chapter2.xlsx.

Database	IMPPAT 1.0	Phytochemica [36]	Polur <i>et al.</i> [44]
Basic statistics			
Number of Indian medicinal plants	1742	5	295
Number of phytochemicals	9596	963	1829
Type of associations			
Plant-phytochemical associations	Yes	Yes	Yes
Plant-therapeutic use associations	Yes	No	Yes
Plant-medicinal formulation associations	Yes	No	No
Phytochemical-human target protein associations	Yes	No	Yes
Plant part-phytochemical associations	No	Yes	No
Additional Features			
Web interface	Yes	Yes	No
Availability of 2D structure of phytochemicals	Yes	No	No
Availability of 3D structure of phytochemicals	Yes	Yes	No
Downloadable structure file formats	MOL, MOL2, SDF, PDB & PDBQT	MOL2	No
Chemical classification	Yes	Yes	No
Physicochemical properties	Yes	Yes	No
ADMET properties	Yes	Yes	No
Drug-likeness scores	Yes	No	No
Cytoscape network visualization of associations	Yes	No	No
Filter phytochemicals based on physicochemical properties	Yes	Yes	No
Filter phytochemicals based on drug-likeness scores	Yes	No	No
Chemical similarity search within database	Yes	No	No

Table 2.1: Comparison of IMPPAT 1.0 with earlier databases on phytochemical composition of Indian medicinal plants.

Feature	IMPPAT 2.0	IMPPAT 1.0
Number of Indian medicinal plants	4,010	1,742
Number of Phytochemicals	17,967	9,596
Number of Plant - Part - Phytochemical associations	189,386	Not available
Number of Plant - Phytochemical associations	124,995	27,074
Number of Therapeutic uses	1,095	1,124
Number of Plant - Part - Therapeutic use associations	89,733	Not available
Number of Plant - Therapeutic use associations	60,732	11,514
Number of Traditional medicinal formulations	1,133	974
Number of Plant - Part – Traditional medicinal formulation associations	7,815	Not available
Number of Plant - Traditional medicinal formulation associations	6,317	5,069

Table 2.2: Comparison of the updated version IMPPAT 2.0 with the previous version 1.0.

Chapter 3

Exploration of the phytochemical space of Indian medicinal plants

In this chapter, we present the results from an in-depth analysis of the enhanced and expanded phytochemical atlas of Indian medicinal plants compiled in IMPPAT 2.0. A primary objective of IMPPAT is to exhaustively capture the phytochemical space of Indian medicinal plants. Thus, through extensive manual curation and standardization, IMPPAT 2.0 provides a FAIR [60] compliant non-redundant stereo-aware library of 17967 phytochemicals with 2D and 3D chemical structures. In order to analyze the phytochemical space of Indian medicinal plants, we characterized the molecular complexity and the molecular scaffold based structural diversity of the phytochemical library of IMPPAT 2.0, and thereafter, compared with other chemical libraries. We then filtered a subset of 1335 drug-like phytochemicals using multiple drug-likeness rules. Finally, we compared the phytochemicals in IMPPAT 2.0 with phytochemicals from Chinese medicinal plants. From our cheminformatics analysis, we find that phytochemicals in IMPPAT 2.0 are more likely enriched with specific protein binders rather than promiscuous binders, have scaffold diversity similar to many larger natural product libraries, and share minimum overlap with the phytochemical space of Chinese medicinal plants. These results highlight the uniqueness, utility and complementary nature of the phytochemical space of

Indian medicinal plants captured in IMPPAT 2.0. The work reported in this chapter is contained in the published manuscript [50] and the manuscript [51].

3.1 Molecular complexity comparison with other collections of small molecules

Small molecules which are selective and specific binders of a target protein are preferable for drug development over promiscuous binders which 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 [52, 144]. In particular, Clemons *et al.* [84] have shown that stereochemical complexity and shape complexity are excellent indicators of target protein specificity of small molecules. The stereochemical complexity is the fraction of stereogenic carbon atoms in a compound, whereas the shape complexity 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.

In their work, Clemons *et al.* [84] 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). The CC has 6152 representative small molecules from commercial sources, DC' has 5963 small molecules synthesized by academic community using methods like diversity-oriented synthesis and NP has 2477 small molecules from natural products. Clemons *et al.* [84] found that CC, DC' and NP molecules on an average have low, intermediate and high values, respectively of both stereochemical and shape complexity. Thereafter, Clemons *et al.* [84] 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, while in comparison DC' molecules with intermediate complexity and NP molecules with high complexity are more enriched in specific binders and depleted in promiscuous binders.

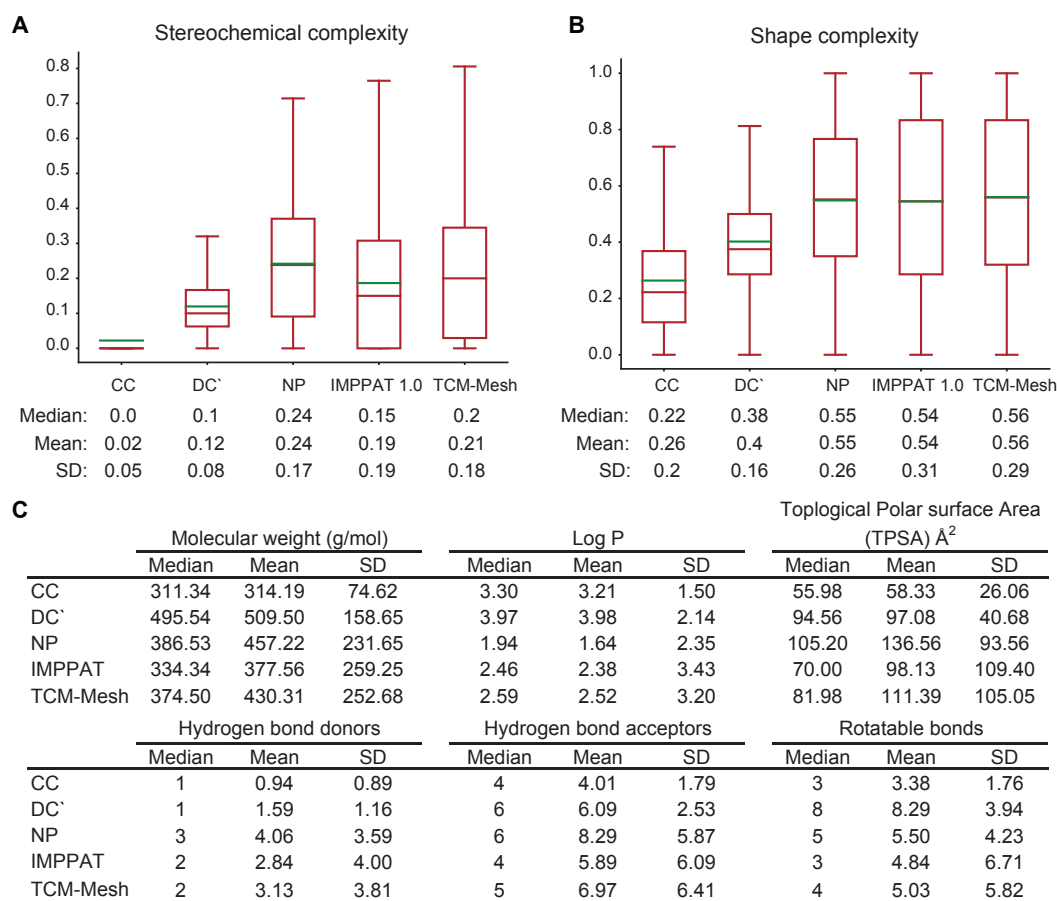


Figure 3.1: Comparison of the molecular complexity of IMPPAT 1.0 with other chemical libraries. (A) Distribution of the stereochemical complexity, and (B) the shape complexity for small molecules in five chemical libraries, namely, CC, DC', NP, IMPPAT 1.0 phytochemicals and TCM-Mesh phytochemicals. Note that the lower end of the box plot is the first quartile, upper end is the third quartile, brown line inside the box is the median, green line is the mean of the distribution. Also, the median, mean and standard deviation (SD) of the distribution is shown below the box plot. (C) Median, mean and SD of six physicochemical properties, namely, molecular weight (g/mol), log P, topological polar surface area (TPSA) (Å²), number of hydrogen bond donors, number of hydrogen bond acceptors and number of rotatable bonds for the chemical libraries CC, DC', NP, IMPPAT 1.0 phytochemicals and TCM-Mesh phytochemicals.

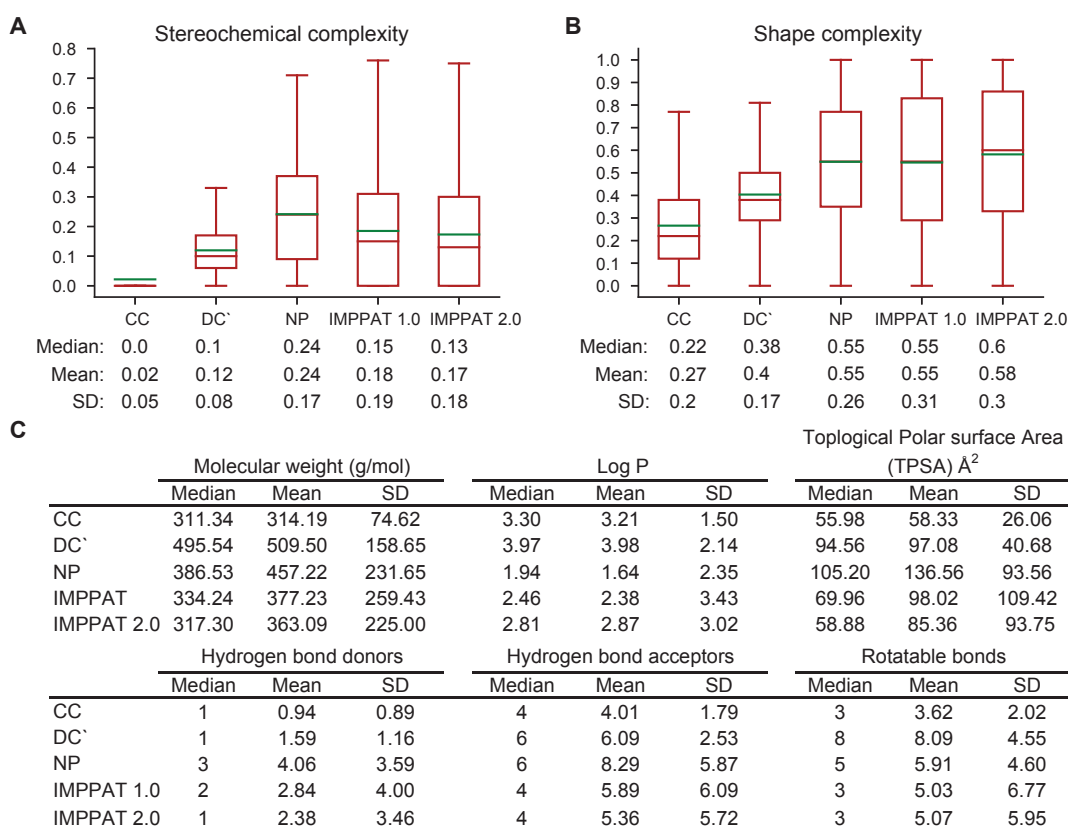


Figure 3.2: Comparison of the molecular complexity of chemical libraries. (A) Distribution of the stereochemical complexity, and (B) the shape complexity for small molecules in five chemical libraries, namely, CC, DC', NP, IMPPAT 1.0 and IMPPAT 2.0. Note that the lower end of the box plot is the first quartile, upper end is the third quartile, brown line inside the box is the median, green line is the mean of the distribution. Also, the median, mean and standard deviation (SD) of the distribution is 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) (Å²), number of hydrogen bond donors, number of hydrogen bond acceptors, and number of rotatable bonds, for small molecules in five chemical libraries.

Lastly, NP molecules were found to be more depleted in promiscuous binders in comparison to DC' molecules [84].

Previously [50], 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 10140 phytochemicals in TCM-Mesh [43] 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 (Figure 3.1). Due to 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 computed using RDKit [74] (Figure 3.2A,B). 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 3.2A,B).

In another study, Clemons *et al.* [145] 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 are very similar to IMPPAT 1.0, and closer to NP and DC' rather than CC collection (Figure 3.2C).

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, a valuable space for ongoing efforts in drug discovery.

3.2 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 [146] of chemical libraries. One way to define the molecular scaffold is via the core structure of a molecule with all its ring system and all chain fragments connecting the rings [78, 146]. Previously, Lipkus *et al.* [136, 137] have analyzed the scaffold diversity of organic compounds compiled in 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.* [136, 137] to compute the molecular scaffold at three levels, namely, graph/node/bond (G/N/B) level, graph/node (G/N) level and graph level using RDKit [74]. Scaffold at G/N/B level has connectivity, element and bond information, at G/N level has connectivity and element information but ignores bond information, and at graph level has only connectivity information [136, 137]. Among the phytochemicals in IMPPAT 2.0, we find 5179 scaffolds at G/N/B level, 4072 at G/N level and 3434 at graph level.

Thereafter, we compared the scaffold diversity of IMPPAT 2.0 with seven other natural product libraries (CMAUP [147], COCONUT [148], NANPDB [149], NPATLAS [49], SuperNatural-II [150], TCM-Mesh [43] and UNPD), approved drugs obtained from Drugbank [45], and more than 100 million organic compounds from PubChem [134] (Table 3.1). 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 3.1).

Figure 3.3A,B show the distribution of the number of rings and number of heteroatoms across the 5179 scaffolds at G/N/B level found in phytochemicals of IMPPAT

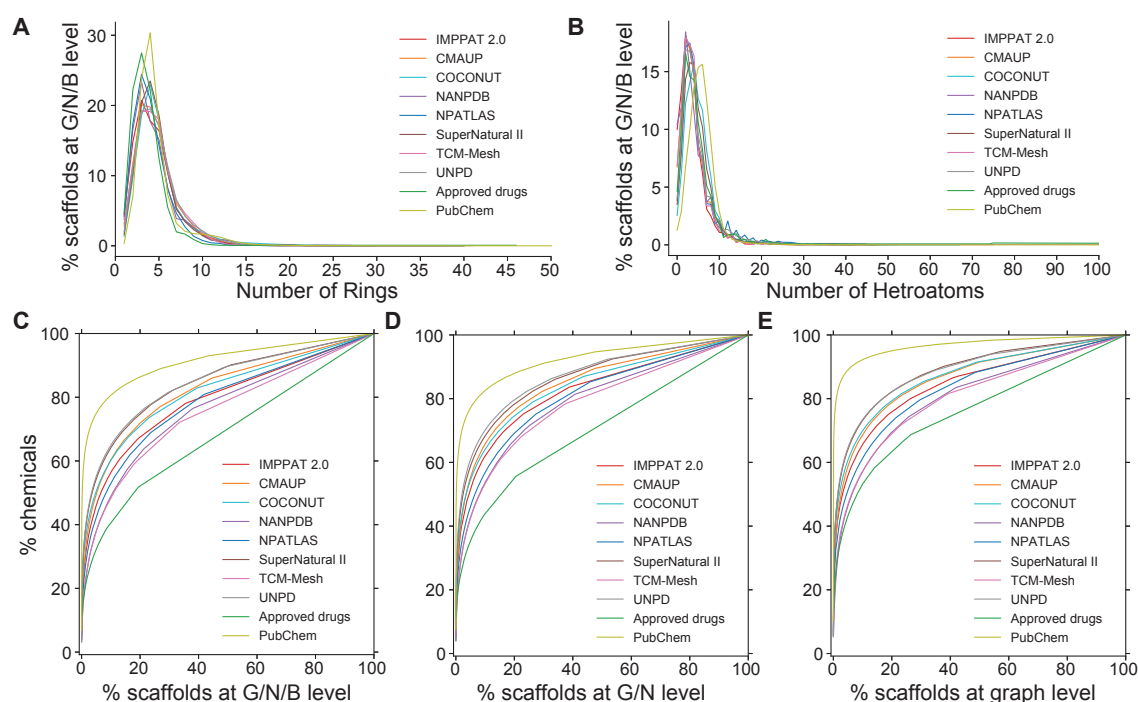


Figure 3.3: Analysis of the scaffold diversity of phytochemicals in IMPPAT 2.0 with seven other natural product libraries, approved drugs, and organic compounds from PubChem [134]. Distribution of (A) the number of ring systems and (B) the number of heteroatoms, in scaffolds at graph/node/bond (G/N/B) level. Cyclic system retrieval (CSR) curves for scaffolds at: (C) G/N/B level, (D) graph/node (G/N) level, and (E) graph level.

2.0. While 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 3.3A). Notably, 231 scaffolds (4.5%) are single ring system, and this indicates high degree of ring diversity in phytochemicals of IMPPAT 2.0. We also find 49.7% of the 5179 scaffolds have two or three or four heteroatoms, and only 0.4% of the scaffolds contain ≥ 20 heteroatoms (Figure 3.3B). Further, 518 scaffolds (10%) are completely composed of carbon atoms. Figure 3.3A,B also show that the distributions of number of rings and number of heteroatoms in scaffolds found in phytochemicals of IMPPAT 2.0 are similar to 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 [136, 137, 151, 152] were plotted for scaffolds computed at G/N/B level (Figure 3.3C), G/N level (Figure 3.3D) and graph level (Figure 3.3E). CSR curves were generated by plotting the percent of scaffolds on the x-axis and the percent of 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 [136, 137, 151–153]. In an ideal distribution with maximum scaffold diversity wherein each compound has a unique scaffold, the CSR curve will be the diagonal line with 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 levels off (Figures 3.3C-E). As we move from scaffolds at G/N/B level (least abstraction) to G/N level to graph level (high abstraction), the scaffold diversity reduces across all the chemical libraries, with CSR curves shifting up away from the diagonal (Figures 3.3C-E).

Importantly, the scaffold diversity of phytochemicals in IMPPAT 2.0 (red) and other natural product libraries lie in between the scaffold diversity of 100 million organic compounds from PubChem (low diversity) and approved drugs (high diversity) (Figure 3.3C-E). Table 3.1 lists the AUC and P_{50} from CSR curves of scaffolds at G/N/B level for the

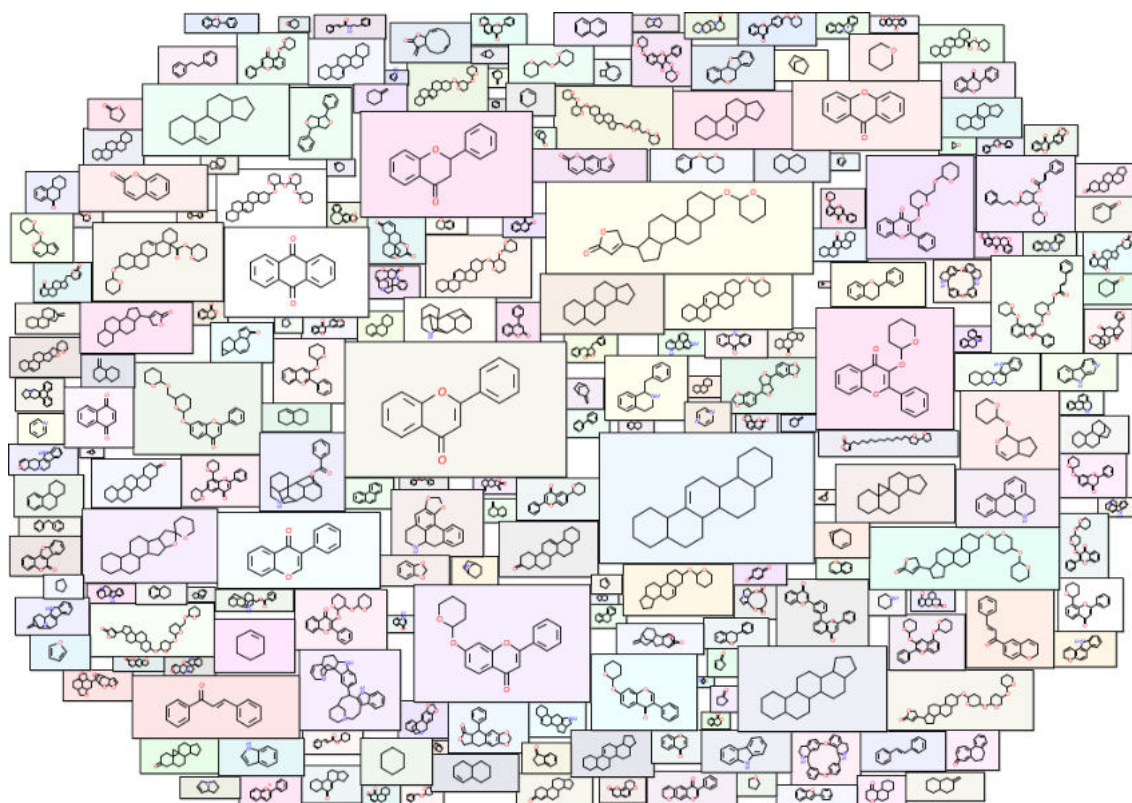


Figure 3.4: Molecular cloud visualization [154, 155] of the top scaffolds at G/N/B level present in phytochemicals of IMPPAT 2.0. The top constitute the 217 scaffolds at G/N/B level that are present in ≥ 10 phytochemicals in IMPPAT 2.0. In this figure, 216 of these top scaffolds are shown after excluding 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.

phytochemicals in IMPPAT 2.0 and other chemical libraries. In line with expectation, the approved drug library was found to be most diverse with AUC of 0.69 and P₅₀ of 17.93% (Table 3.1). Interestingly, the scaffold diversity of phytochemicals in IMPPAT 2.0 was found to be greater than 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 P₅₀ of 6.58%, respectively (Table 3.1). Further, 64.5% of the 5179 scaffolds at G/N/B level found in phytochemicals of IMPPAT 2.0 are singletons which are present in only one compound (Table 3.1). In contrast, 217 scaffolds present in 10 or more phytochemicals cumulatively account for 43.6% of the phytochemicals in IMPPAT 2.0, and a molecule cloud visualization [154,155] of these scaffolds is shown in Figure 3.4 (after excluding 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.

3.3 Drug-like phytochemical space

Natural products have been an important source of approved drugs [17,18]. To predict the subset of drug-like phytochemicals in IMPPAT 2.0, we used six scoring schemes namely, Lipinski's rule of five (RO5) [156], Ghose rule [157], Veber rule [158], Egan rule [159], Pfizer 3/75 rule [160] and GlaxoSmithKline's (GSK) 4/400 rule [161]. RO5 [156] is a classical rule of thumb to filter drug-like small molecules based on four physiochemical properties. RO5 considers a small molecule to be drug-like if at least 3 of the following criteria are met: hydrogen bond donors ≤ 5 , hydrogen bond acceptors ≤ 10 , molecular weight < 500 g/mol and $\log P \leq 5$ [156]. Ghose rule considers a small molecule drug-like if $\log P$ is ≥ -0.4 and ≤ 5.6 , molecular weight is ≥ 160 g/mol and ≤ 480 g/mol, molar refractivity is ≥ 40 and ≤ 130 , and total number of atoms is ≥ 20 and ≤ 70 [157]. Veber rule considers small molecules to have good oral bioavailability, and hence more, drug-like if number of rotatable bonds ≤ 10 and either topological polar surface area (TPSA)

$\leq 140 \text{ \AA}^2$ or total hydrogen bond donors and acceptors ≤ 12 [158]. Egan rule considers small molecules to have good oral bioavailability, and hence, more drug-like if their log P is between -1.0 and 6, and TPSA is $\leq 132 \text{ \AA}^2$ [159]. Pfizer 3/75 rule considers small molecules with $\log P < 3$ and $\text{TPSA} > 75 \text{ \AA}^2$ to be likely less toxic, and hence, more drug-like [160]. GSK 4/400 rule considers a small molecule more drug-like if it has both molecular weight $< 400 \text{ g/mol}$ and $\log P < 4$ [161].

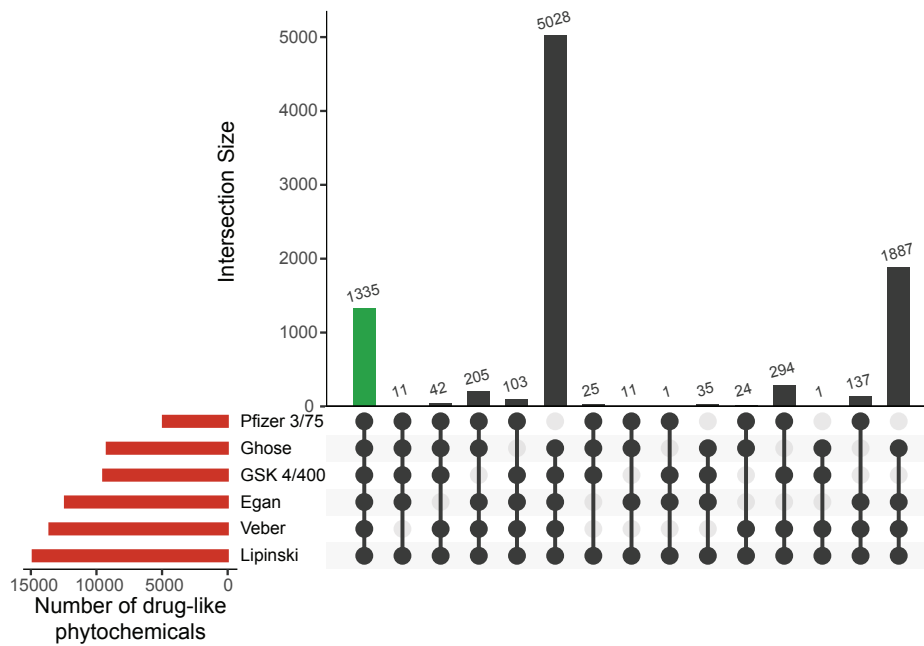
Figure 3.5A is an UpSet [162] visualization of the set intersections of phytochemicals that pass one or more of these six rules. Majority of the phytochemicals pass RO5 (14847), followed by Veber (13574) and Egan (12390) rules. Pfizer 3/75 was found to be the most restrictive rule, with 4924 phytochemicals passing it. A drug-like subset of 1335 phytochemicals is identified based on the stringent criteria of passing all six rules (Figure 3.5A; Supplementary Table S3.1).

The top 5 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 3.5B shows the chemical classification for the 1335 drug-like phytochemicals obtained using ClassyFire [79]. The top 3 chemical super-classes namely, Phenylpropanoids and polyketides, Lipids and lipid-like molecules, and Organoheterocyclic compounds account for 486, 253, and 245 drug-like phytochemicals, respectively.

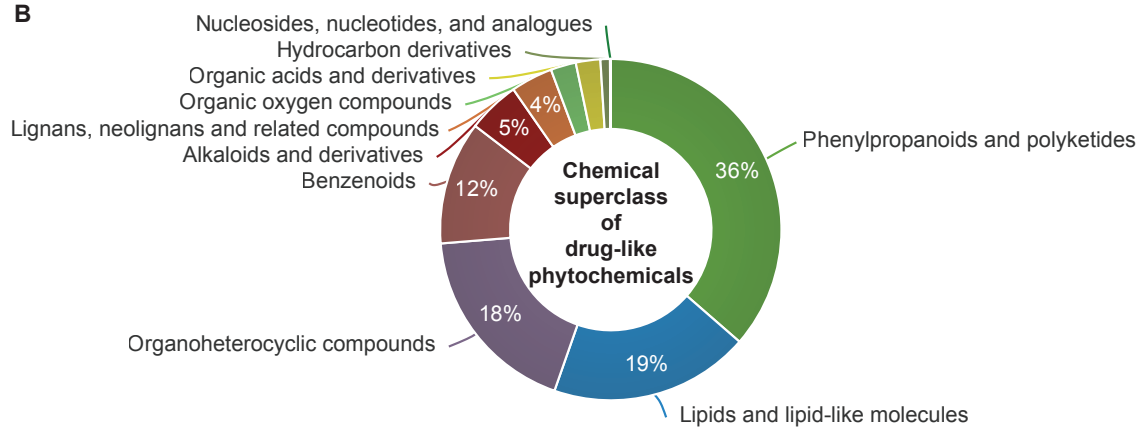
Weighted quantitative estimate of drug-likeness (QEDw) score can also be used to assess drug-likeness of small molecules, and this measure can take values between 0 (least drug-like) to 1 (most drug-like) [163]. For the 1335 drug-like phytochemicals, Figure 3.5C shows the distribution of QEDw scores 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 ≥ 0.80 .

We also compared the 1335 drug-like phytochemicals in IMPPAT 2.0 with the drugs approved by United States Federal Drug Administration (US FDA). Chemical structure

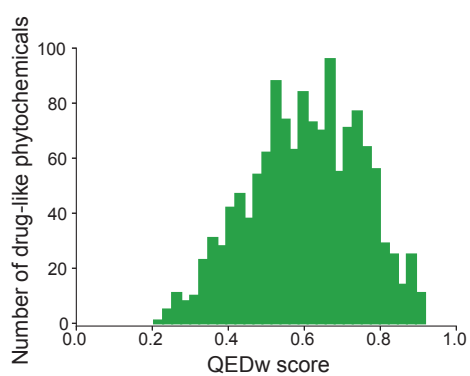
A



B



C



D

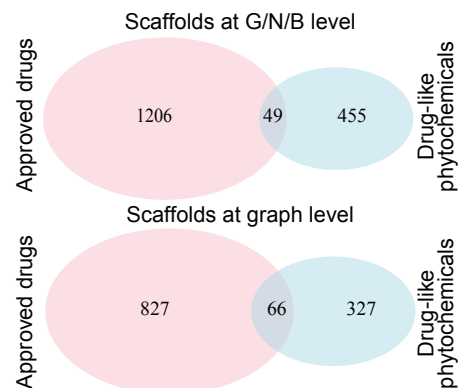


Figure 3.5 (previous page): Drug-likeness analysis of phytochemicals in IMPPAT 2.0. (A) UpSet plot visualization of the set intersections of phytochemicals that pass one or more of the six drug-likeness rules. The horizontal bars show the number of phytochemicals which pass the different drug-likeness rules. The vertical bars show the set intersections between phytochemicals that pass different drug-likeness rules. The green bar shows the 1335 phytochemicals which pass all six drug-likeness rules. This plot was generated using UpSetR package [162]. (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.

similarity between any two molecules is quantified using the widely-used metric, Tanimoto coefficient (T_c) [83] which was computed using Extended Circular Fingerprints (ECFP4) as implemented in RDKit [74, 164]. A set of 2567 approved drugs were obtained from DrugBank [45] version 5.1.9. Based on chemical similarity ($T_c \geq 0.50$), 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, while 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 G/N/B, G/N and graph levels, respectively (Figure 3.5D). Thus, the drug-like phytochemicals in IMPPAT 2.0 presents a unique chemical scaffold space with minimal overlap with approved drugs. These results highlight the potential of our database 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. Figure 3.6A shows the drug-like CSN wherein nodes correspond to phytochemicals and an edge exists between any pair of phytochemicals if $T_c \geq 0.5$. The value of T_c 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. The drug-like CSN is very sparse with graph density of 0.01,

and it can be partitioned into 90 connected components (with at least 2 nodes each) and 210 isolated nodes. In Figure 3.6A, 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 2 phytochemicals (Colchicine and its metabolite Colchicine) are approved drugs and remaining phytochemicals are similar to them. For each of the top 12 components, the maximum common substructure (MCS) is shown in Figure 3.6B. The MCS for phytochemicals in a connected component of the CSN was computed using FindMCS function in RDKit [74]. The substructures confirm the structural uniqueness of the different connected components. In sum, the CSN highlights the chemical dissimilarity, and hence, the structural diversity of the drug-like space of 1335 phytochemicals.

3.4 Comparison with the phytochemical space of Chinese medicinal plants

Previously [50], a comparison of the 9596 phytochemicals in IMPPAT 1.0 with the 10140 phytochemicals in TCM-Mesh [43] revealed that less than 25% phytochemicals (2305) in IMPPAT 1.0 are present in the TCM-Mesh. Notably, TCM-Mesh is a large-scale database compiling information on 10140 phytochemicals produced by 6235 Chinese medicinal plants [43]. We also performed a comparison of the 17967 phytochemicals in IMPPAT 2.0 with the 10140 phytochemicals in TCM-Mesh. Though 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 which is shared with TCM-Mesh has decreased to 18.6% (Figure 3.7A).

Further, we compared the drug-like subset of 1335 phytochemicals in IMPPAT 2.0 with the corresponding drug-like subset in TCM-Mesh. Specifically, a subset of 938 drug-like phytochemicals was obtained in TCM-Mesh based on the six rules (Figure 3.7B). Further, Figure 3.7C shows the distribution of QEDw scores for the 938 drug-like phy-

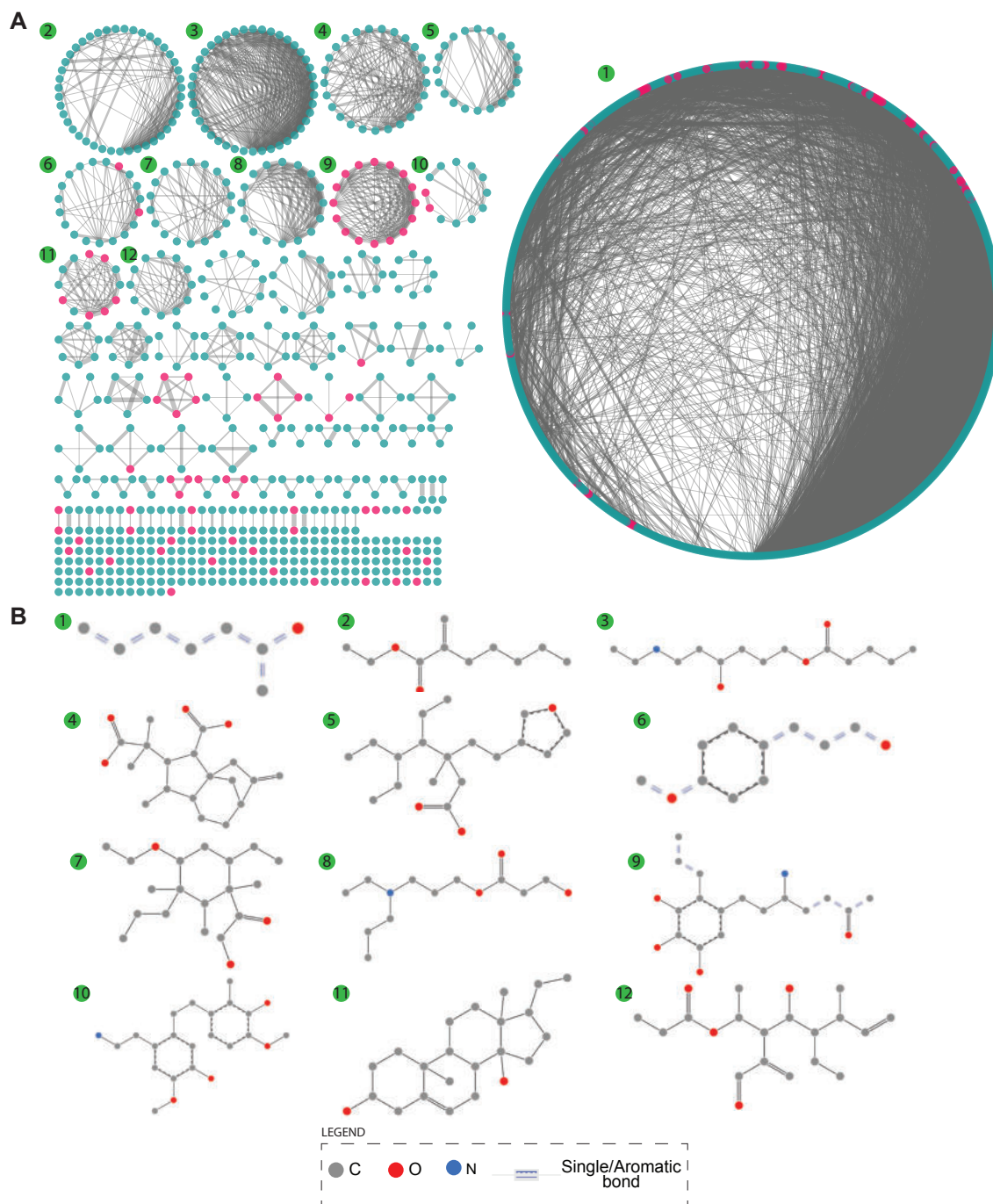


Figure 3.6: (A) Chemical similarity network (CSN) of the 1335 drug-like phytochemicals in IMPPAT 2.0. The degree sorted circle layout in Cytoscape [165] is used to visualize the CSN. Cyan nodes correspond to drug-like phytochemicals that are not similar to any approved drug and pink nodes to those that are similar to at least one approved drugs. 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 SMARTSview web-server [166, 167].

tochemicals 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.

3.5 Discussion

The cheminformatics analysis of the phytochemicals in IMPPAT 2.0 revealed that their stereochemical complexity and shape complexity is similar to 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. Also, we find that the scaffold diversity of natural product libraries including IMPPAT 2.0 lies in between the scaffold diversity of more than 100 million organic compounds from PubChem (low diversity) and approved drugs (high diversity). This highlights the utility of our phytochemical library for the identification of biologically active new chemical entities with novel scaffolds. Using six drug-likeness scores, we identified a subset of 1335 drug-like phytochemicals which pass all six rules considered here. We find that only 11 of the drug-like phytochemicals are already approved drugs. Also, the drug-like phytochemicals and approved drugs have very few common scaffolds, revealing the pool of scaffolds present in drug-like phytochemicals in IMPPAT 2.0 but not present in 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

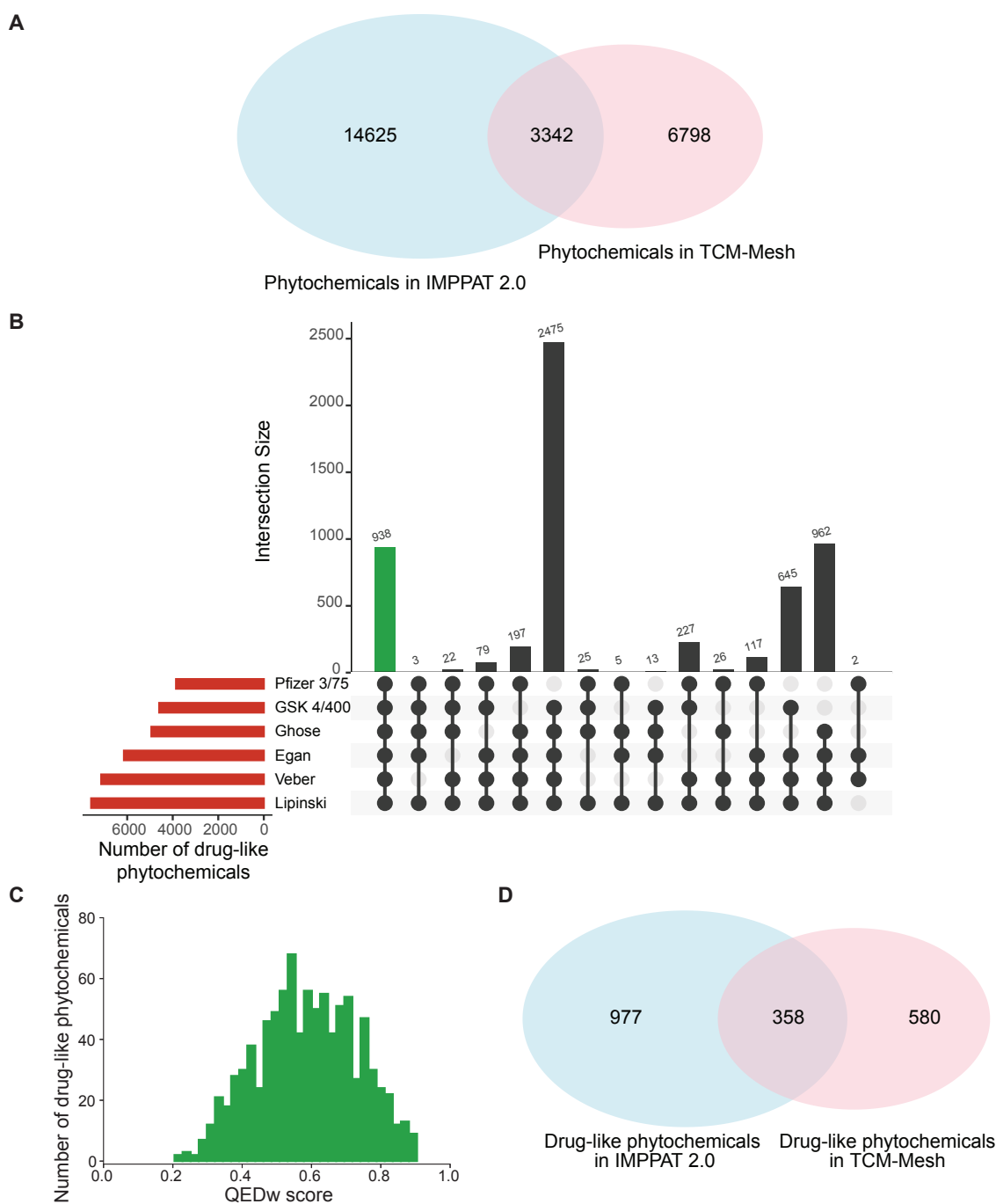


Figure 3.7: Comparison of the phytochemical space of Indian medicinal plants and Chinese medicinal plants. (A) Venn diagram shows the overlap between the phytochemicals in IMPPAT 2.0 and TCM-Mesh. (B) UpSet plot visualization of the set intersections of phytochemicals in TCM-Mesh that pass one or more of the six drug-likeness rules. The horizontal bars show the number of phytochemicals which 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 which 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.

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.

Supplementary Information

Supplementary Table S3.1 associated with this chapter is available for download from the GitHub repository: https://github.com/asamallab/PhDThesis-Vivek_Ananth_RP/blob/main/SI/ST_Chapter3.xlsx.

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	15226	5179	3338	0.34	0.22	0.64	0.79	6.58
NPATLAS	31099	10227	5947	0.33	0.19	0.58	0.79	8.35
COCONUT	385926	109024	65963	0.28	0.17	0.61	0.82	4.82
CMAUP	43987	11105	6151	0.25	0.14	0.55	0.82	5.15
UNPD	215585	44281	22514	0.21	0.1	0.51	0.85	3.39
SuperNatural II	308998	62125	30453	0.2	0.1	0.49	0.85	3.61
PubChem	101452728	12493379	7059386	0.12	0.07	0.57	0.91	0.22

Table 3.1: Scaffold diversity of phytochemicals in IMPPAT 2.0, and comparison with other chemical libraries. The molecular scaffolds are computed at graph/node/bond (G/N/B) level. Here, M is 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.

Chapter 4

Compilation, curation and exploration of a chemical atlas of secondary metabolites from medicinal fungi

In this chapter, we present our research on building the natural product space 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 [25, 66–69]. Medicinal mushrooms [25, 66–69] have been used for centuries in traditional medicine, especially traditional Chinese medicine, to treat human ailments. Presently, the valuable information on secondary metabolites and therapeutic uses of medicinal fungi is dispersed across published literature including articles and books [25, 66–69], and this limits its effective use for drug discovery. A common repository of high-quality information on secondary metabolites and therapeutic uses of medicinal fungi is thus needed to harness the potential of this natural product space for drug discovery.

In this chapter we describe our manually curated database namely, Medicinal Fungi

Secondary metabolites And Therapeutics (MeFSAT), dedicated to secondary metabolites and therapeutic uses of medicinal fungi which is openly accessible at: <https://cb.imsc.res.in/mefsat/>. The work reported in this chapter is contained in the published manuscript [52].

4.1 Workflow for the compilation and curation of MeFSAT database

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 (Figure 4.1).

4.1.1 Compilation of 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 PubMed [168] 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 (Supplementary Table S4.1). Apart from research articles, we also curated information on medicinal fungi from books [25, 66–69] 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 Checklist [169] and Mycobank [170]. In the end, this mapping to accepted names led to a unique list of 253 medicinal fungi (Supplementary Table S4.2). Note that the curated list of 253 medicinal fungi does not include Lichens [171].

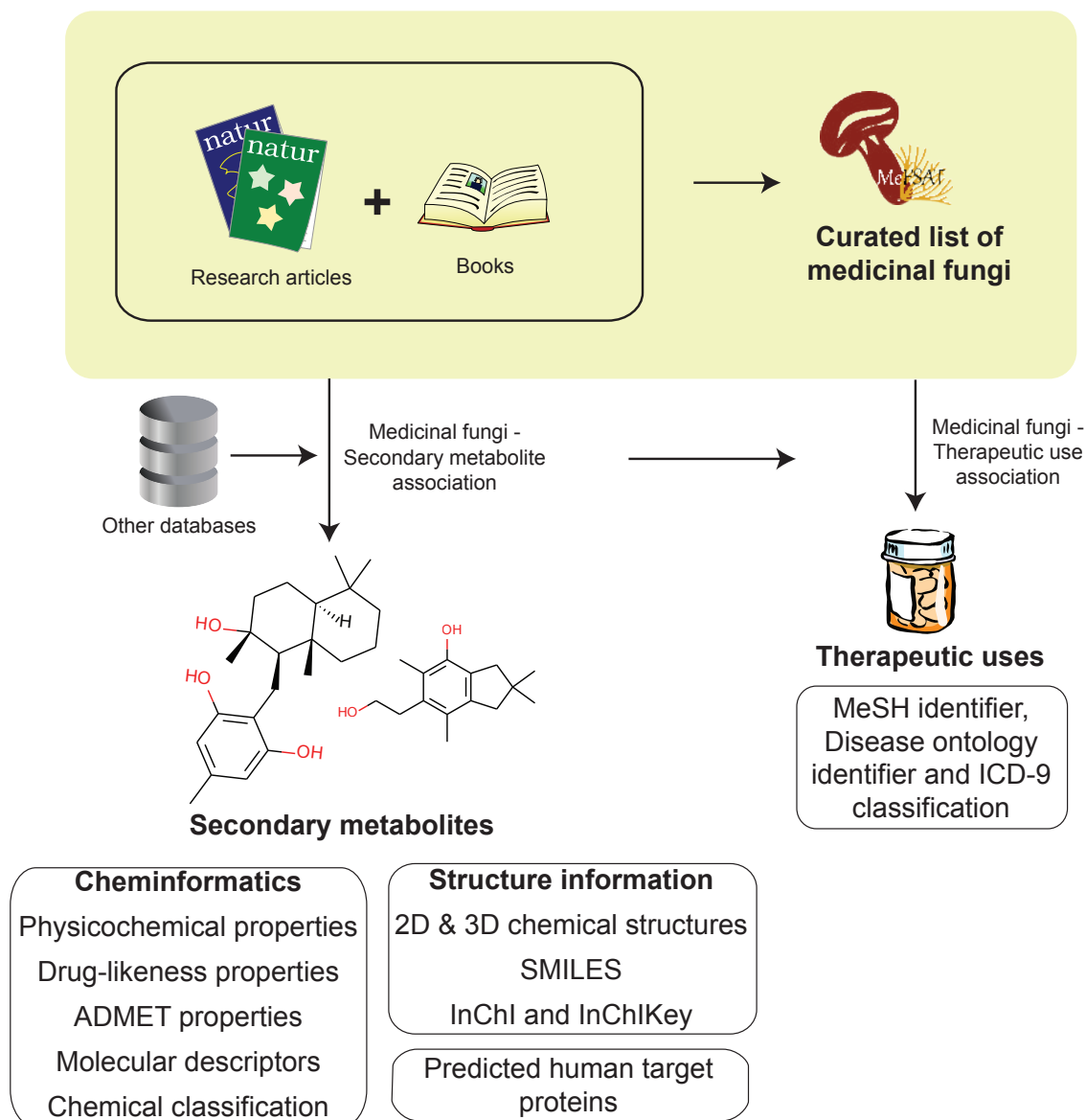


Figure 4.1: Schematic overview of the workflow to construct the MeFSAT database. Briefly, we compiled a curated list of medicinal fungi from the published literature. Next, we mined the published literature to compile secondary metabolites produced by different medicinal fungi. For the manually curated non-redundant list of secondary metabolites produced by medicinal fungi, we compiled the chemical structures and employed cheminformatics tools to compute their physicochemical, drug-likeness and ADMET properties. Subsequently, we compiled and curated information on the therapeutic uses of medicinal fungi from the published literature.

4.1.2 Compilation of the 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 (Supplementary Table S4.2). From research articles and books [25, 66–69, 172], 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, NPATLAS [48] 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.

4.1.3 Curated *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 [105], ChemIDplus (<https://chem.nlm.nih.gov/chemidplus/>), Chempider [173] and NPATLAS [48]. 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 coefficient [83] (T_c) to determine chemical similarity between secondary metabolites. To create a non-redundant chemical library, we merged compiled secondary metabolites from published literature if they were determined to be identical based on their chemical structure. Importantly, our manual curation effort 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 (Supplementary Tables S4.3 and S4.4).

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.

4.1.4 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 OpenBabel [74] while their 3D chemical structures were manually retrieved from PubChem [105] if available. For the remaining secondary metabolites whose 3D structures are not available in PubChem, the 3D structures were generated using RDKit [74] (<http://www.rdkit.org/>) by embedding the molecule using ETKDG method [174] followed by energy minimization using MMFF94 force field [175]. The 3D structures for secondary metabolites were saved in SDF, MOL, MOL2, PDB and PDBQT file formats using OpenBabel [75]. Apart

from the 2D and 3D structure information, the SMILES, InChI and InChIKey of the secondary metabolites were also generated using OpenBabel [75] (Supplementary Table S4.5). Moreover, the secondary metabolites in MeFSAT were hierarchically classified into chemical kingdom, chemical superclass, chemical class and chemical subclass using ClassyFire [79] (<http://classyfire.wishartlab.com/>) (Supplementary Table S4.5).

The basic physicochemical properties of the secondary metabolites in MeFSAT database were computed using RDKit [74] and SwissADME [135] (<http://www.swissadme.ch/>) (Supplementary Table S4.6). To assess the drug-likeness of the secondary metabolites in MeFSAT, we computed multiple scoring schemes and properties namely, Lipinski's rule of five (RO5) [176], Ghose filter [157], Veber filter [158], Egan filter [159], Pfizer 3/75 filter [160], GlaxoSmithKline's (GSK) 4/400 [161], number of leadlikeness violations [177] and weighted quantitative estimate of drug-likeness (QEDw) [163] using RDKit [74] and SwissADME [135] (Supplementary Table S4.7). The assessment of Absorption, Distribution, Metabolism, Excretion and Toxicity (ADMET) properties is essential in the drug discovery pipeline. We used SwissADME [135] to predict the ADMET properties of the secondary metabolites in MeFSAT (Supplementary Table S4.8). Finally, we computed 1875 (2D and 3D) molecular descriptors for each secondary metabolite in MeFSAT using PaDEL [142] (<http://www.yapcsoft.com/dd/padeldescriptor/>). These molecular descriptors can be broadly categorized into different classes such as chemical composition, topology, 3D shape and functionality.

4.1.5 Genome sequences of medicinal fungi

For the 184 medicinal fungi that have secondary metabolite information in MeFSAT, we gathered their genome sequencing status from the Joint Genome Institute (JGI) portal [178] (<https://genome.jgi.doe.gov/>) and the National Center for Biotechnology Information (NCBI) [168] (<https://www.ncbi.nlm.nih.gov/genome>). In addition, we compiled information from published literature on the traditional system of medicine

in which these medicinal fungi are used.


4.1.6 Compilation and curation of 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 [25,69]. Note that the compiled therapeutic uses are based on available information on the use of medicinal fungi to treat human diseases. Notably, we manually curated the compiled therapeutic use terms from various literature sources to create a unique list of standardized therapeutic use terms for medicinal fungi in MeFSAT by mapping the therapeutic use terms to standard terms from Medical Subject Headings (MeSH) [128], Disease Ontology [125] and ICD-9-CM chapters [179]. 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 (Supplementary Table S4.9). Note that MeFSAT compiles therapeutic uses at the level of medicinal fungi rather than secondary metabolites from published literature.

4.1.7 Predicted human target proteins of secondary metabolites

Lastly, we compiled the predicted human target proteins of secondary metabolites in MeFSAT from the STITCH [55] database (<http://stitch.embl.de/>). To date, STITCH [55] is the largest resource on predicted interactions between chemicals and their target proteins. In MeFSAT database, we included only high confidence interactions between secondary metabolites and human target proteins that have a combined STITCH [55] score ≥ 700 . Further, we also mapped the genes corresponding to predicted human target proteins of secondary metabolites from STITCH [55] to their respective HUGO Gene Nomenclature Committee (HGNC) symbols [180].

A Boletus edulis



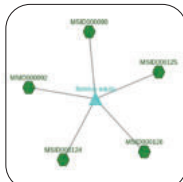
Family: Boletaceae

System of Traditional medicine: Traditional Chinese Medicine

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

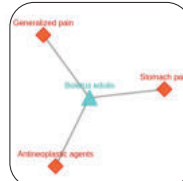
Genome sequence

Medicinal Fungi	Secondary metabolite identifier	Secondary metabolite name	Reference
Boletus edulis	MSD000090	[5S]-[3-ethenylphenyl]-1,2-ethanedithiol	PMID:17440932
Boletus edulis	MSD000092	[5S]-[4-acetylphenyl]-1,2-ethanedithiol	PMID:17440932
Boletus edulis	MSD000124	1-(3-Ethylphenyl)-1,2-ethanedithiol	PMID:17440932
Boletus edulis	MSD000125	1-(3-Ferrocenylphenyl)-ethanone	PMID:17440932
Boletus edulis	MSD000126	1-(4-Ethylphenyl)-1,2-ethanedithiol	PMID:17440932



C

Medicinal fungi	Therapeutic use	Therapeutic use identifier	Reference
Boletus edulis	Antiviral agents	MESH:D000970, MESH:D009369, ICD-9:159.1, ICD-9:239.8, ICD-9:E533.1, DDD:162	PMID:30668382
Boletus edulis	Generalized pain	ICD-9:780.36	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's Rule Violation | GSK 4400

Pfizer 375 | Veber Rule | Egan Rule

QEDw

Search

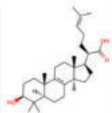
Physicochemical filter | Drug-like filter | Chemical similarity filter

Enter SMILES | Choose Fingerprint

Search

B Secondary metabolite: Trametenolic acid B

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



Molecular formula: C₂₃H₄₀O₃

SMILES: CC1=CC(=C(C=C1)C(=O)O)C2=CC(=C(C=C2)C(=O)O)C3=CC(=C(C=C3)C(=O)O)C4=CC(=C(C=C4)C(=O)O)C5=CC(=C(C=C5)C(=O)O)C6=CC(=C(C=C6)C(=O)O)C7=CC(=C(C=C7)C(=O)O)C8=CC(=C(C=C8)C(=O)O)C9=CC(=C(C=C9)C(=O)O)C10=CC(=C(C=C10)C(=O)O)C11=CC(=C(C=C11)C(=O)O)C12=CC(=C(C=C12)C(=O)O)C13=CC(=C(C=C13)C(=O)O)C14=CC(=C(C=C14)C(=O)O)C15=CC(=C(C=C15)C(=O)O)C16=CC(=C(C=C16)C(=O)O)C17=CC(=C(C=C17)C(=O)O)C18=CC(=C(C=C18)C(=O)O)C19=CC(=C(C=C19)C(=O)O)C20=CC(=C(C=C20)C(=O)O)C21=CC(=C(C=C21)C(=O)O)C22=CC(=C(C=C22)C(=O)O)C23=CC(=C(C=C23)C(=O)O)C24=CC(=C(C=C24)C(=O)O)C25=CC(=C(C=C25)C(=O)O)C26=CC(=C(C=C26)C(=O)O)C27=CC(=C(C=C27)C(=O)O)C28=CC(=C(C=C28)C(=O)O)C29=CC(=C(C=C29)C(=O)O)C30=CC(=C(C=C30)C(=O)O)C31=CC(=C(C=C31)C(=O)O)C32=CC(=C(C=C32)C(=O)O)C33=CC(=C(C=C33)C(=O)O)C34=CC(=C(C=C34)C(=O)O)C35=CC(=C(C=C35)C(=O)O)C36=CC(=C(C=C36)C(=O)O)C37=CC(=C(C=C37)C(=O)O)C38=CC(=C(C=C38)C(=O)O)C39=CC(=C(C=C39)C(=O)O)C40=CC(=C(C=C40)C(=O)O)C41=CC(=C(C=C41)C(=O)O)C42=CC(=C(C=C42)C(=O)O)C43=CC(=C(C=C43)C(=O)O)C44=CC(=C(C=C44)C(=O)O)C45=CC(=C(C=C45)C(=O)O)C46=CC(=C(C=C46)C(=O)O)C47=CC(=C(C=C47)C(=O)O)C48=CC(=C(C=C48)C(=O)O)C49=CC(=C(C=C49)C(=O)O)C50=CC(=C(C=C50)C(=O)O)C51=CC(=C(C=C51)C(=O)O)C52=CC(=C(C=C52)C(=O)O)C53=CC(=C(C=C53)C(=O)O)C54=CC(=C(C=C54)C(=O)O)C55=CC(=C(C=C55)C(=O)O)C56=CC(=C(C=C56)C(=O)O)C57=CC(=C(C=C57)C(=O)O)C58=CC(=C(C=C58)C(=O)O)C59=CC(=C(C=C59)C(=O)O)C60=CC(=C(C=C60)C(=O)O)C61=CC(=C(C=C61)C(=O)O)C62=CC(=C(C=C62)C(=O)O)C63=CC(=C(C=C63)C(=O)O)C64=CC(=C(C=C64)C(=O)O)C65=CC(=C(C=C65)C(=O)O)C66=CC(=C(C=C66)C(=O)O)C67=CC(=C(C=C67)C(=O)O)C68=CC(=C(C=C68)C(=O)O)C69=CC(=C(C=C69)C(=O)O)C70=CC(=C(C=C70)C(=O)O)C71=CC(=C(C=C71)C(=O)O)C72=CC(=C(C=C72)C(=O)O)C73=CC(=C(C=C73)C(=O)O)C74=CC(=C(C=C74)C(=O)O)C75=CC(=C(C=C75)C(=O)O)C76=CC(=C(C=C76)C(=O)O)C77=CC(=C(C=C77)C(=O)O)C78=CC(=C(C=C78)C(=O)O)C79=CC(=C(C=C79)C(=O)O)C80=CC(=C(C=C80)C(=O)O)C81=CC(=C(C=C81)C(=O)O)C82=CC(=C(C=C82)C(=O)O)C83=CC(=C(C=C83)C(=O)O)C84=CC(=C(C=C84)C(=O)O)C85=CC(=C(C=C85)C(=O)O)C86=CC(=C(C=C86)C(=O)O)C87=CC(=C(C=C87)C(=O)O)C88=CC(=C(C=C88)C(=O)O)C89=CC(=C(C=C89)C(=O)O)C90=CC(=C(C=C90)C(=O)O)C91=CC(=C(C=C91)C(=O)O)C92=CC(=C(C=C92)C(=O)O)C93=CC(=C(C=C93)C(=O)O)C94=CC(=C(C=C94)C(=O)O)C95=CC(=C(C=C95)C(=O)O)C96=CC(=C(C=C96)C(=O)O)C97=CC(=C(C=C97)C(=O)O)C98=CC(=C(C=C98)C(=O)O)C99=CC(=C(C=C99)C(=O)O)C100=CC(=C(C=C100)C(=O)O)C101=CC(=C(C=C101)C(=O)O)C102=CC(=C(C=C102)C(=O)O)C103=CC(=C(C=C103)C(=O)O)C104=CC(=C(C=C104)C(=O)O)C105=CC(=C(C=C105)C(=O)O)C106=CC(=C(C=C106)C(=O)O)C107=CC(=C(C=C107)C(=O)O)C108=CC(=C(C=C108)C(=O)O)C109=CC(=C(C=C109)C(=O)O)C110=CC(=C(C=C110)C(=O)O)C111=CC(=C(C=C111)C(=O)O)C112=CC(=C(C=C112)C(=O)O)C113=CC(=C(C=C113)C(=O)O)C114=CC(=C(C=C114)C(=O)O)C115=CC(=C(C=C115)C(=O)O)C116=CC(=C(C=C116)C(=O)O)C117=CC(=C(C=C117)C(=O)O)C118=CC(=C(C=C118)C(=O)O)C119=CC(=C(C=C119)C(=O)O)C120=CC(=C(C=C120)C(=O)O)C121=CC(=C(C=C121)C(=O)O)C122=CC(=C(C=C122)C(=O)O)C123=CC(=C(C=C123)C(=O)O)C124=CC(=C(C=C124)C(=O)O)C125=CC(=C(C=C125)C(=O)O)C126=CC(=C(C=C126)C(=O)O)C127=CC(=C(C=C127)C(=O)O)C128=CC(=C(C=C128)C(=O)O)C129=CC(=C(C=C129)C(=O)O)C130=CC(=C(C=C130)C(=O)O)C131=CC(=C(C=C131)C(=O)O)C132=CC(=C(C=C132)C(=O)O)C133=CC(=C(C=C133)C(=O)O)C134=CC(=C(C=C134)C(=O)O)C135=CC(=C(C=C135)C(=O)O)C136=CC(=C(C=C136)C(=O)O)C137=CC(=C(C=C137)C(=O)O)C138=CC(=C(C=C138)C(=O)O)C139=CC(=C(C=C139)C(=O)O)C140=CC(=C(C=C140)C(=O)O)C141=CC(=C(C=C141)C(=O)O)C142=CC(=C(C=C142)C(=O)O)C143=CC(=C(C=C143)C(=O)O)C144=CC(=C(C=C144)C(=O)O)C145=CC(=C(C=C145)C(=O)O)C146=CC(=C(C=C146)C(=O)O)C147=CC(=C(C=C147)C(=O)O)C148=CC(=C(C=C148)C(=O)O)C149=CC(=C(C=C149)C(=O)O)C150=CC(=C(C=C150)C(=O)O)C151=CC(=C(C=C151)C(=O)O)C152=CC(=C(C=C152)C(=O)O)C153=CC(=C(C=C153)C(=O)O)C154=CC(=C(C=C154)C(=O)O)C155=CC(=C(C=C155)C(=O)O)C156=CC(=C(C=C156)C(=O)O)C157=CC(=C(C=C157)C(=O)O)C158=CC(=C(C=C158)C(=O)O)C159=CC(=C(C=C159)C(=O)O)C160=CC(=C(C=C160)C(=O)O)C161=CC(=C(C=C161)C(=O)O)C162=CC(=C(C=C162)C(=O)O)C163=CC(=C(C=C163)C(=O)O)C164=CC(=C(C=C164)C(=O)O)C165=CC(=C(C=C165)C(=O)O)C166=CC(=C(C=C166)C(=O)O)C167=CC(=C(C=C167)C(=O)O)C168=CC(=C(C=C168)C(=O)O)C169=CC(=C(C=C169)C(=O)O)C170=CC(=C(C=C170)C(=O)O)C171=CC(=C(C=C171)C(=O)O)C172=CC(=C(C=C172)C(=O)O)C173=CC(=C(C=C173)C(=O)O)C174=CC(=C(C=C174)C(=O)O)C175=CC(=C(C=C175)C(=O)O)C176=CC(=C(C=C176)C(=O)O)C177=CC(=C(C=C177)C(=O)O)C178=CC(=C(C=C178)C(=O)O)C179=CC(=C(C=C179)C(=O)O)C180=CC(=C(C=C180)C(=O)O)C181=CC(=C(C=C181)C(=O)O)C182=CC(=C(C=C182)C(=O)O)C183=CC(=C(C=C183)C(=O)O)C184=CC(=C(C=C184)C(=O)O)C185=CC(=C(C=C185)C(=O)O)C186=CC(=C(C=C186)C(=O)O)C187=CC(=C(C=C187)C(=O)O)C188=CC(=C(C=C188)C(=O)O)C189=CC(=C(C=C189)C(=O)O)C190=CC(=C(C=C190)C(=O)O)C191=CC(=C(C=C191)C(=O)O)C192=CC(=C(C=C192)C(=O)O)C193=CC(=C(C=C193)C(=O)O)C194=CC(=C(C=C194)C(=O)O)C195=CC(=C(C=C195)C(=O)O)C196=CC(=C(C=C196)C(=O)O)C197=CC(=C(C=C197)C(=O)O)C198=CC(=C(C=C198)C(=O)O)C199=CC(=C(C=C199)C(=O)O)C200=CC(=C(C=C200)C(=O)O)C201=CC(=C(C=C201)C(=O)O)C202=CC(=C(C=C202)C(=O)O)C203=CC(=C(C=C203)C(=O)O)C204=CC(=C(C=C204)C(=O)O)C205=CC(=C(C=C205)C(=O)O)C206=CC(=C(C=C206)C(=O)O)C207=CC(=C(C=C207)C(=O)O)C208=CC(=C(C=C208)C(=O)O)C209=CC(=C(C=C209)C(=O)O)C210=CC(=C(C=C210)C(=O)O)C211=CC(=C(C=C211)C(=O)O)C212=CC(=C(C=C212)C(=O)O)C213=CC(=C(C=C213)C(=O)O)C214=CC(=C(C=C214)C(=O)O)C215=CC(=C(C=C215)C(=O)O)C216=CC(=C(C=C216)C(=O)O)C217=CC(=C(C=C217)C(=O)O)C218=CC(=C(C=C218)C(=O)O)C219=CC(=C(C=C219)C(=O)O)C220=CC(=C(C=C220)C(=O)O)C221=CC(=C(C=C221)C(=O)O)C222=CC(=C(C=C222)C(=O)O)C223=CC(=C(C=C223)C(=O)O)C224=CC(=C(C=C224)C(=O)O)C225=CC(=C(C=C225)C(=O)O)C226=CC(=C(C=C226)C(=O)O)C227=CC(=C(C=C227)C(=O)O)C228=CC(=C(C=C228)C(=O)O)C229=CC(=C(C=C229)C(=O)O)C230=CC(=C(C=C230)C(=O)O)C231=CC(=C(C=C231)C(=O)O)C232=CC(=C(C=C232)C(=O)O)C233=CC(=C(C=C233)C(=O)O)C234=CC(=C(C=C234)C(=O)O)C235=CC(=C(C=C235)C(=O)O)C236=CC(=C(C=C236)C(=O)O)C237=CC(=C(C=C237)C(=O)O)C238=CC(=C(C=C238)C(=O)O)C239=CC(=C(C=C239)C(=O)O)C240=CC(=C(C=C240)C(=O)O)C241=CC(=C(C=C241)C(=O)O)C242=CC(=C(C=C242)C(=O)O)C243=CC(=C(C=C243)C(=O)O)C244=CC(=C(C=C244)C(=O)O)C245=CC(=C(C=C245)C(=O)O)C246=CC(=C(C=C246)C(=O)O)C247=CC(=C(C=C247)C(=O)O)C248=CC(=C(C=C248)C(=O)O)C249=CC(=C(C=C249)C(=O)O)C250=CC(=C(C=C250)C(=O)O)C251=CC(=C(C=C251)C(=O)O)C252=CC(=C(C=C252)C(=O)O)C253=CC(=C(C=C253)C(=O)O)C254=CC(=C(C=C254)C(=O)O)C255=CC(=C(C=C255)C(=O)O)C256=CC(=C(C=C256)C(=O)O)C257=CC(=C(C=C257)C(=O)O)C258=CC(=C(C=C258)C(=O)O)C259=CC(=C(C=C259)C(=O)O)C260=CC(=C(C=C260)C(=O)O)C261=CC(=C(C=C261)C(=O)O)C262=CC(=C(C=C262)C(=O)O)C263=CC(=C(C=C263)C(=O)O)C264=CC(=C(C=C264)C(=O)O)C265=CC(=C(C=C265)C(=O)O)C266=CC(=C(C=C266)C(=O)O)C267=CC(=C(C=C267)C(=O)O)C268=CC(=C(C=C268)C(=O)O)C269=CC(=C(C=C269)C(=O)O)C270=CC(=C(C=C270)C(=O)O)C271=CC(=C(C=C271)C(=O)O)C272=CC(=C(C=C272)C(=O)O)C273=CC(=C(C=C273)C(=O)O)C274=CC(=C(C=C274)C(=O)O)C275=CC(=C(C=C275)C(=O)O)C276=CC(=C(C=C276)C(=O)O)C277=CC(=C(C=C277)C(=O)O)C278=CC(=C(C=C278)C(=O)O)C279=CC(=C(C=C279)C(=O)O)C280=CC(=C(C=C280)C(=O)O)C281=CC(=C(C=C281)C(=O)O)C282=CC(=C(C=C282)C(=O)O)C283=CC(=C(C=C283)C(=O)O)C284=CC(=C(C=C284)C(=O)O)C285=CC(=C(C=C285)C(=O)O)C286=CC(=C(C=C286)C(=O)O)C287=CC(=C(C=C287)C(=O)O)C288=CC(=C(C=C288)C(=O)O)C289=CC(=C(C=C289)C(=O)O)C290=CC(=C(C=C290)C(=O)O)C291=CC(=C(C=C291)C(=O)O)C292=CC(=C(C=C292)C(=O)O)C293=CC(=C(C=C293)C(=O)O)C294=CC(=C(C=C294)C(=O)O)C295=CC(=C(C=C295)C(=O)O)C296=CC(=C(C=C296)C(=O)O)C297=CC(=C(C=C297)C(=O)O)C298=CC(=C(C=C298)C(=O)O)C299=CC(=C(C=C299)C(=O)O)C300=CC(=C(C=C300)C(=O)O)C301=CC(=C(C=C301)C(=O)O)C302=CC(=C(C=C302)C(=O)O)C303=CC(=C(C=C303)C(=O)O)C304=CC(=C(C=C304)C(=O)O)C305=CC(=C(C=C305)C(=O)O)C306=CC(=C(C=C306)C(=O)O)C307=CC(=C(C=C307)C(=O)O)C308=CC(=C(C=C308)C(=O)O)C309=CC(=C(C=C309)C(=O)O)C310=CC(=C(C=C310)C(=O)O)C311=CC(=C(C=C311)C(=O)O)C312=CC(=C(C=C312)C(=O)O)C313=CC(=C(C=C313)C(=O)O)C314=CC(=C(C=C314)C(=O)O)C315=CC(=C(C=C315)C(=O)O)C316=CC(=C(C=C316)C(=O)O)C317=CC(=C(C=C317)C(=O)O)C318=CC(=C(C=C318)C(=O)O)C319=CC(=C(C=C319)C(=O)O)C320=CC(=C(C=C320)C(=O)O)C321=CC(=C(C=C321)C(=O)O)C322=CC(=C(C=C322)C(=O)O)C323=CC(=C(C=C323)C(=O)O)C324=CC(=C(C=C324)C(=O)O)C325=CC(=C(C=C325)C(=O)O)C326=CC(=C(C=C326)C(=O)O)C327=CC(=C(C=C327)C(=O)O)C328=CC(=C(C=C328)C(=O)O)C329=CC(=C(C=C329)C(=O)O)C330=CC(=C(C=C330)C(=O)O)C331=CC(=C(C=C331)C(=O)O)C332=CC(=C(C=C332)C(=O)O)C333=CC(=C(C=C333)C(=O)O)C334=CC(=C(C=C334)C(=O)O)C335=CC(=C(C=C335)C(=O)O)C336=CC(=C(C=C336)C(=O)O)C337=CC(=C(C=C337)C(=O)O)C338=CC(=C(C=C338)C(=O)O)C339=CC(=C(C=C339)C(=O)O)C340=CC(=C(C=C340)C(=O)O)C341=CC(=C(C=C341)C(=O)O)C342=CC(=C(C=C342)C(=O)O)C343=CC(=C(C=C343)C(=O)O)C344=CC(=C(C=C344)C(=O)O)C345=CC(=C(C=C345)C(=O)O)C346=CC(=C(C=C346)C(=O)O)C347=CC(=C(C=C347)C(=O)O)C348=CC(=C(C=C348)C(=O)O)C349=CC(=C(C=C349)C(=O)O)C350=CC(=C(C=C350)C(=O)O)C351=CC(=C(C=C351)C(=O)O)C352=CC(=C(C=C352)C(=O)O)C353=CC(=C(C=C353)C(=O)O)C354=CC(=C(C=C354)C(=O)O)C355=CC(=C(C=C355)C(=O)O)C356=CC(=C(C=C356)C(=O)O)C357=CC(=C(C=C357)C(=O)O)C358=CC(=C(C=C358)C(=O)O)C359=CC(=C(C=C359)C(=O)O)C360=CC(=C(C=C360)C(=O)O)C361=CC(=C(C=C361)C(=O)O)C362=CC(=C(C=C362)C(=O)O)C363=CC(=C(C=C363)C(=O)O)C364=CC(=C(C=C364)C(=O)O)C365=CC(=C(C=C365)C(=O)O)C366=CC(=C(C=C366)C(=O)O)C367=CC(=C(C=C367)C(=O)O)C368=CC(=C(C=C368)C(=O)O)C369=CC(=C(C=C369)C(=O)O)C370=CC(=C(C=C370)C(=O)O)C371=CC(=C(C=C371)C(=O)O)C372=CC(=C(C=C372)C(=O)O)C373=CC(=C(C=C373)C(=O)O)C374=CC(=C(C=C374)C(=O)O)C375=CC(=C(C=C375)C(=O)O)C376=CC(=C(C=C376)C(=O)O)C377=CC(=C(C=C377)C(=O)O)C378=CC(=C(C=C378)C(=O)O)C379=CC(=C(C=C379)C(=O)O)C380=CC(=C(C=C380)C(=O)O)C381=CC(=C(C=C381)C(=O)O)C382=CC(=C(C=C382)C(=O)O)C383=CC(=C(C=C383)C(=O)O)C384=CC(=C(C=C384)C(=O)O)C385=CC(=C(C=C385)C(=O)O)C386=CC(=C(C=C386)C(=O)O)C387=CC(=C(C=C387)C(=O)O)C388=CC(=C(C=C388)C(=O)O)C389=CC(=C(C=C389)C(=O)O)C390=CC(=C(C=C390)C(=O)O)C391=CC(=C(C=C391)C(=O)O)C392=CC(=C(C=C392)C(=O)O)C393=CC(=C(C=C393)C(=O)O)C394=CC(=C(C=C394)C(=O)O)C395=CC(=C(C=C395)C(=O)O)C396=CC(=C(C=C396)C(=O)O)C397=CC(=C(C=C397)C(=O)O)C398=CC(=C(C=C398)C(=O)O)C399=CC(=C(C=C399)C(=O)O)C400=CC(=C(C=C400)C(=O)O)C401=CC(=C(C=C401)C(=O)O)C402=CC(=C(C=C402)C(=O)O)C403=CC(=C(C=C403)C(=O)O)C404=CC(=C(C=C404)C(=O)O)C405=CC(=C(C=C405)C(=O)O)C406=CC(=C(C=C406)C(=O)O)C407=CC(=C(C=C407)C(=O)O)C408=CC(=C(C=C408)C(=O)O)C409=CC(=C(C=C409)C(=O)O)C410=CC(=C(C=C410)C(=O)O)C411=CC(=C(C=C411)C(=O)O)C412=CC(=C(C=C412)C(=O)O)C413=CC(=C(C=C413)C(=O)O)C414=CC(=C(C=C414)C(=O)O)C415=CC(=C(C=C415)C(=O)O)C416=CC(=C(C=C416)C(=O)O)C417=CC(=C(C=C417)C(=O)O)C418=CC(=C(C=C418)C(=O)O)C419=CC(=C(C=C419)C(=O)O)C420=CC(=C(C=C420)C(=O)O)C421=CC(=C(C=C421)C(=O)O)C422=CC(=C(C=C422)C(=O)O)C423=CC(=C(C=C423)C(=O)O)C424=CC(=C(C=C424)C(=O)O)C425=CC(=C(C=C425)C(=O)O)C426=CC(=C(C=C426)C(=O)O)C427=CC(=C(C=C427)C(=O)O)C428=CC(=C(C=C428)C(=O)O)C429=CC(=C(C=C429)C(=O)O)C430=CC(=C(C=C430)C(=O)O)C431=CC(=C(C=C431)C(=O)O)C432=CC(=C(C=C432)C(=O)O)C433=CC(=C(C=C433)C(=O)O)C434=CC(=C(C=C434)C(=O)O)C435=CC(=C(C=C435)C(=O)O)C436=CC(=C(C=C436)C(=O)O)C437=CC(=C(C=C437)C(=O)O)C438=CC(=C(C=C438)C(=O)O)C439=CC(=C(C=C439)C(=O)O)C440=CC(=C(C=C440)C(=O)O)C441=CC(=C(C=C441)C(=O)O)C442=CC(=C(C=C442)C(=O)O)C443=CC(=C(C=C443)C(=O)O)C444=CC(=C(C=C444)C(=O)O)C445=CC(=C(C=C445)C(=O)O)C446=CC(=C(C=C446)C(=O)O)C447=CC(=C(C=C447)C(=O)O)C448=CC(=C(C=C448)C(=O)O)C449=CC(=C(C=C449)C(=O)O)C450=CC(=C(C=C450)C(=O)O)C451=CC(=C(C=C451)C(=O)O)C452=CC(=C(C=C452)C(=O)O)C453=CC(=C(C=C453)C(=O)O)C454=CC(=C(C=C454)C(=O)O)C455=CC(=C(C=C455)C(=O)O)C456=CC(=C(C=C456)C(=O)O)C457=CC(=C(C=C457)C(=O)O)C458=CC(=C(C=C458)C(=O)O)C459=CC(=C(C=C459)C(=O)O)C460=CC(=C(C=C460)C(=O)O)C461=CC(=C(C=C461)C(=O)O)C462=CC(=C(C=C462)C(=O)O)C463=CC(=C(C=C463)C(=O)O)C464=CC(=C(C=C464)C(=O)O)C465=CC(=C(C=C465)C(=O)O)C466=CC(=C(C=C466)C(=O)O)C467=CC(=C(C=C467)C(=O)O)C468=CC(=C(C=C468)C(=O)O)C469=CC(=C(C=C469)C(=O)O)C470=CC(=C(C=C470)C(=O)O)C471=CC(=C(C=C471)C(=O)O)C472=CC(=C(C=C472)C(=O)O)C473=CC(=C(C=C473)C(=O)O)C474=CC(=C(C=C474)C(=O)O)C475=CC(=C(C=C475)C(=O)O)C476=CC(=C(C=C476)C(=O)O)C477=CC(=C(C=C477)C(=O)O)C478=CC(=C(C=C478)C(=O)O)C479=CC(=C(C=C479)C(=O)O)C480=CC(=C(C=C480)C(=O)O)C481=CC(=C(C=C481)C(=O)O)C482=CC(=C(C=C482)C(=O)O)C483=CC(=C(C=C483)C(=O)O)C484=CC(=C(C=C484)C(=O)O)C485=CC(=C(C=C485)C(=O)O)C486=CC(=C(C=C486)C(=O)O)C487=CC(=C(C=C487)C(=O)O)C488=CC(=C(C=C488)C(=O)O)C489=CC(=C(C=C489)C(=O)O)C490=CC(=C(C=C490)C(=O)O)C491=CC(=C(C=C491)C(=O)O)C492=CC(=C(C=C492)C(=O)O)C493=CC(=C(C=C493)C(=O)O)C494=CC(=C(C=C494)C(=O)O)C495=CC(=C(C=C495)C(=O)O)C496=CC(=C(C=C496)C(=O)O)C497=CC(=C(C=C497)C(=O)O)C498=CC(=C(C=C498)C(=O)O)C499=CC(=C(C=C499)C(=O)O)C500=CC(=C(C=C500)C(=O)O)C501=CC(=C(C=C501)C(=O)O)C502=CC(=C(C=C502)C(=O)O)C503=CC(=C(C=C503)C(=O)O)C504=CC(=C(C=C504)C(=O)O)C505=CC(=C(C=C505)C(=O)O)C506=CC(=C(C=C506)C(=O)O)C507=CC(=C(C=C507)C(=O)O)C508=CC(=C(C=C508)C(=O)O)C509=CC(=C(C=C509)C(=O)O)C510=CC

Figure 4.2 (previous page): 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. 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.

4.2 Web-interface of MeFSAT

MeFSAT provides the compiled information on 184 medicinal fungi, 1830 secondary metabolites and 149 therapeutic uses via a modern and intuitive web-interface (Figure 4.2) openly accessible at: <https://cb.imsc.res.in/mefsat/>. It has been created using a similar approach used for IMPPAT web-interface. The compiled information in MeFSAT is stored in a SQL database created using the open-source relational database management system MariaDB (<https://mariadb.org/>). To render visualizations in the web-interface, we used Cytoscape.js (<http://js.cytoscape.org/>) and Google Charts (<https://developers.google.com/chart>).

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 (Figure 4.2). The query result is displayed as a table of associations with relevant literature references. In the 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 (Figure 4.2A). 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, chemical structure in different file formats, physicochemical properties, drug-likeness properties, predicted ADMET properties, molecular descriptors and predicted human target proteins (Figure 4.2B).

In the 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 [128], disease ontology [125] and ICD-9-CM chapters [179] (Figure 4.2C). 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 (Figure 4.2D). 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.

4.3 Exploration of the 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 (Supplementary Tables S4.3–S4.8). Interestingly, we find 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 (Supplementary Table S4.3).

There are 2127 medicinal fungi – secondary metabolite associations in MeFSAT database which encompass 184 medicinal fungi and 1830 secondary metabolites (Figure 4.1; Supplementary Table S4.4). These 1830 secondary metabolites are distributed across 13 chemical superclasses as computed using ClassyFire [79] (Figure 4.3A and Supplementary Table S4.5). Notably, more than 60% of the 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 ‘Benzenoids’ (7%) (Figure 4.3A). 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 Figure 4.3B, 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

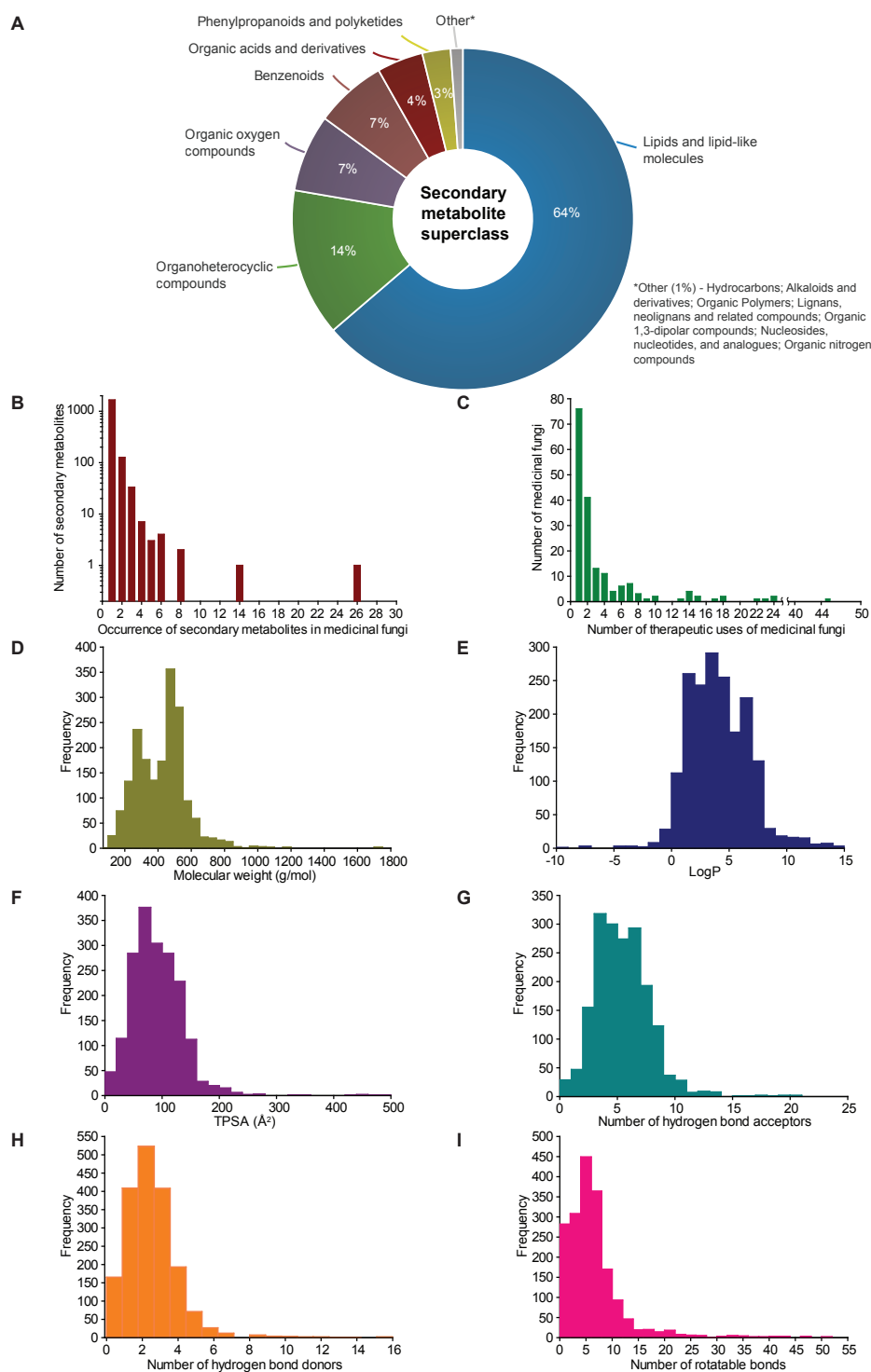


Figure 4.3: Basic statistics for medicinal fungi, their secondary metabolites and therapeutic uses in MeFSAT database. (A) Pie chart shows the distribution of the secondary metabolites in MeFSAT across chemical superclasses obtained from ClassyFire. (B) Histogram of the number of medicinal fungi with 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), 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.

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 STITCH [55] 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 (Figure 4.1 and Supplementary Table S4.9). Figure 4.3C 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 linteus* (22).

4.4 Comparison of the molecular complexity of secondary metabolites in MeFSAT with other small molecule collections

For the 1830 secondary metabolites in MeFSAT database, we computed several physicochemical properties (Supplementary Table S4.6). In Figure 4.3D–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.

We compared the secondary metabolites in MeFSAT with three other small molecule

collections studied by Clemons *et al.* [84]. 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.* [84] and secondary metabolites in our database. Clemons *et al.* [84] have shown that two size-independent molecular complexity 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 [84]. Specifically, Clemons *et al.* [84] 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. 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 [84]. In other words, natural products [50,84] were found to have higher stereochemical and shape complexity while being specific binders of target proteins.

Here, we 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 (Figure 4.4 A,B). Interestingly, we find 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 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 (Figure 4.4C).

4.5 Drug-like secondary metabolites of medicinal fungi

Natural products have directly or indirectly contributed to the discovery of ~34% of the small molecule drugs approved by the US FDA [17, 18]. We used six scoring schemes namely, Lipinski's RO5 [176], Ghose filter [157], Veber filter [158], Egan filter [159], Pfizer 3/75 filter [160] and GSK 4/400 [161] to assess the drug-likeness of 1830 secondary metabolites in MeFSAT. Notably, we identified a subset of 228 secondary metabolites (~12%) in MeFSAT to be drug-like, and these metabolites have passed all the six scoring schemes mentioned above. In Figure 4.5A, 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 [181]. Therefore, we expect a higher fraction of secondary metabolites in MeFSAT, greater than the ~12% metabolites that pass the six scoring schemes, have the potential to be developed into drugs. Figure 4.5B 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 QED_w metric, we find that 19 out of 228 drug-like secondary metabolites in MeFSAT have a high QED_w value of >0.80 (Figure 4.5C). Moreover, Figure 4.6 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 (~34%) pass at least 5 out of

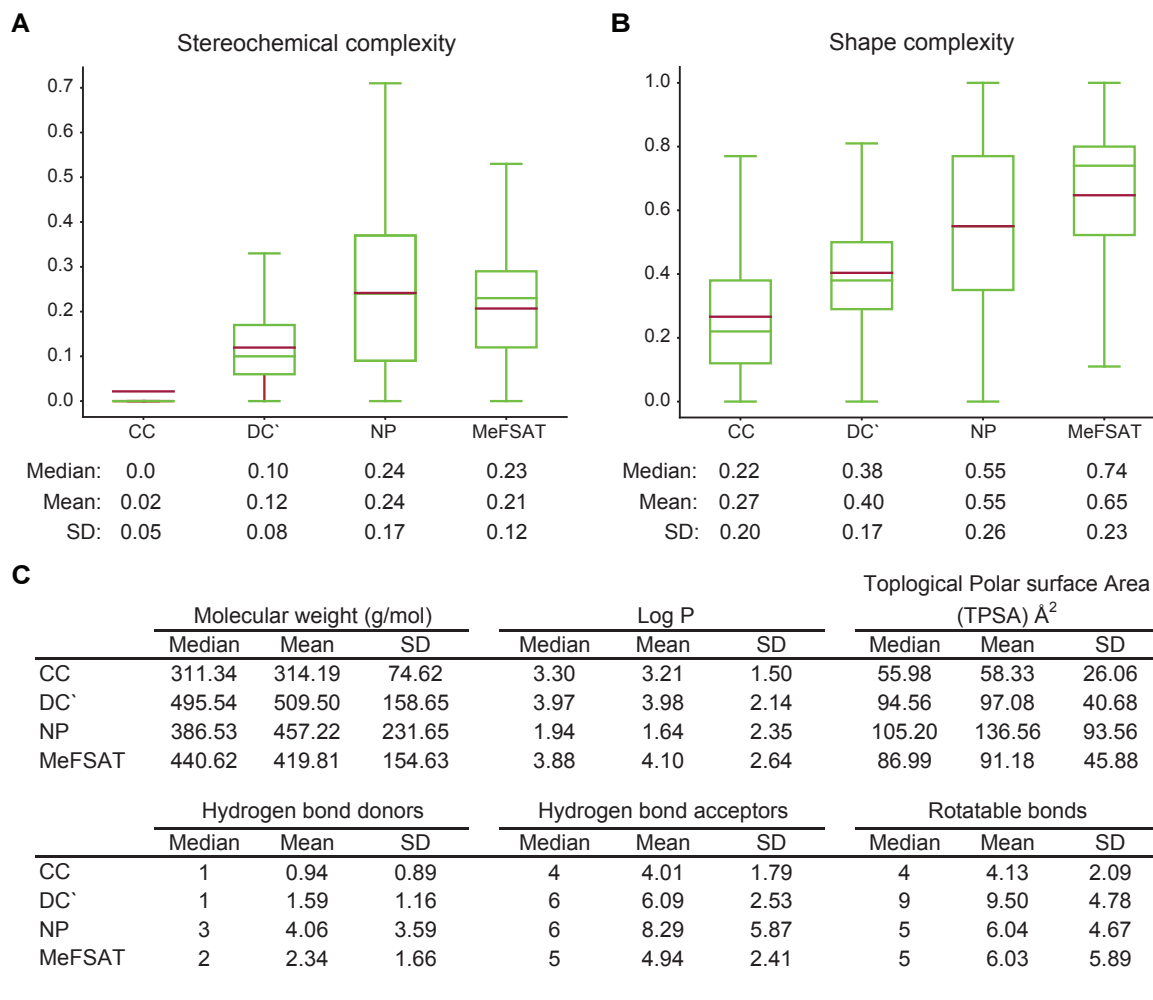


Figure 4.4: Comparison of the stereochemical complexity, shape complexity and physicochemical properties of secondary metabolites in MeFSAT with other small molecule collections. Box plot shows the (A) distribution of the stereochemical complexity and (B) shape complexity of the small molecule collections CC, DC', NP and MeFSAT secondary metabolites. The median, mean and standard deviation (SD) of the distribution 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 donors, number of hydrogen bond acceptors, and number of rotatable bonds for the small molecule collections CC, DC', NP, and MeFSAT secondary metabolites.

the 6 scoring schemes (Figure 4.6). This further highlights the potential of the MeFSAT chemical space for drug discovery.

4.6 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. Tanimoto coefficient [83] (T_c) is a widely used metric to compute chemical structure similarity [182], and we have used T_c based on Extended Circular Fingerprints (ECFP4) [164] 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 [45]. 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.* [183], we choose the cutoff of $T_c \geq 0.5$ to decide if a given pair of chemicals have significant structure similarity. We constructed chemical similarity networks (CSNs) wherein nodes are chemicals and edges between pairs of chemicals signify high chemical similarity. In the CSN, we only retain edges between two chemicals if the T_c between them is ≥ 0.5 .

We constructed two CSNs corresponding to the 1830 secondary metabolites (Figure 4.7) in MeFSAT and the subset of 228 drug-like secondary metabolites (Figure 4.8A) in MeFSAT. Further, these CSNs were visualized using Cytoscape [165]. 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 (Figure 4.7 and Figure 4.8A). The thickness 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

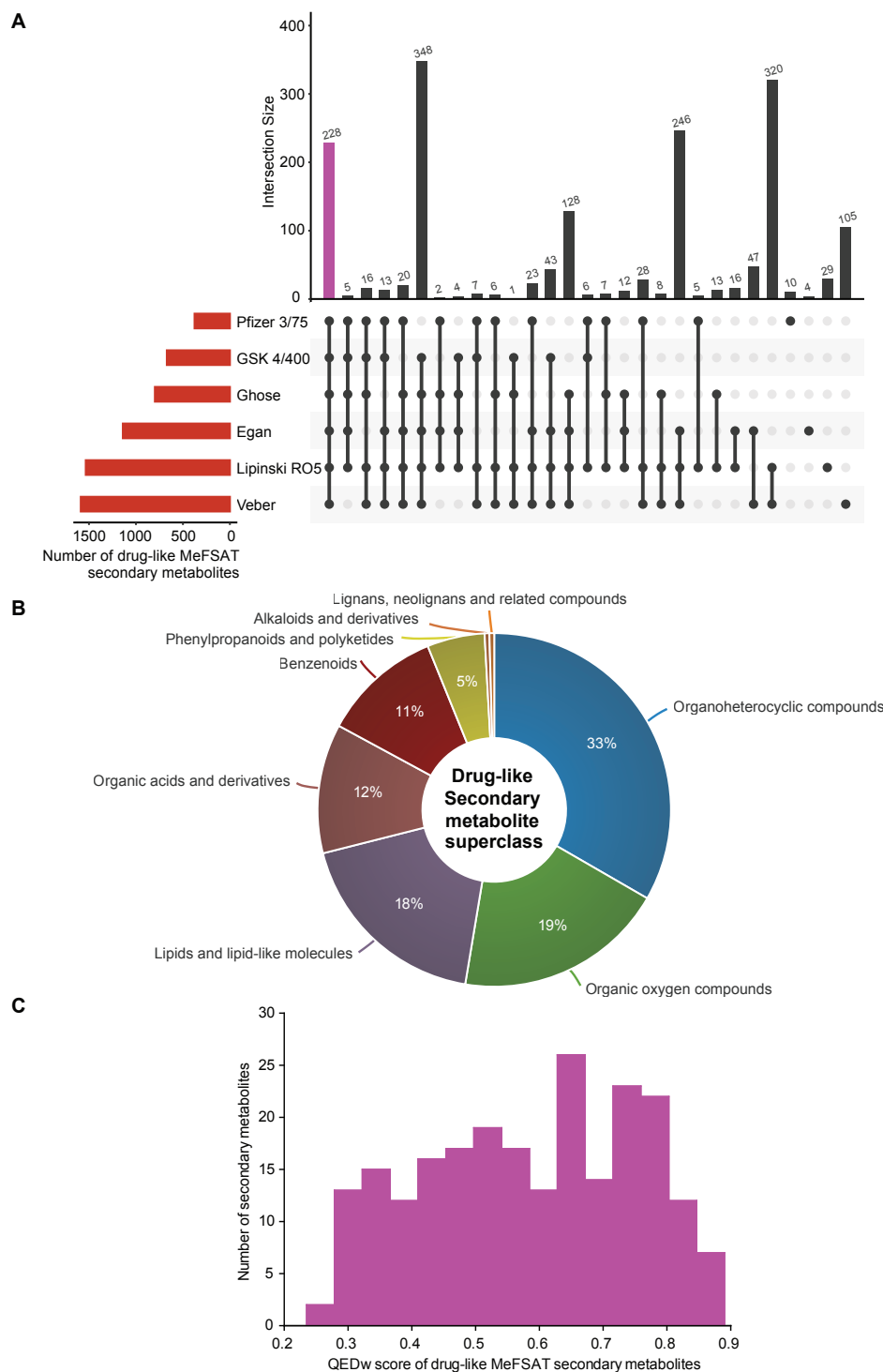


Figure 4.5: Drug-likeness analysis of the secondary metabolites in MeFSAT database. (A) Evaluation of drug-likeness of secondary metabolites based on multiple scores. The horizontal bar plot shows the number of secondary metabolites in MeFSAT database that satisfy different drug-likeness scoring schemes. 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. (B) Classification of the 228 drug-like secondary metabolites into chemical superclasses obtained from ClassyFire. (C) Distribution of the QEDw scores for the 228 drug-like secondary metabolites.

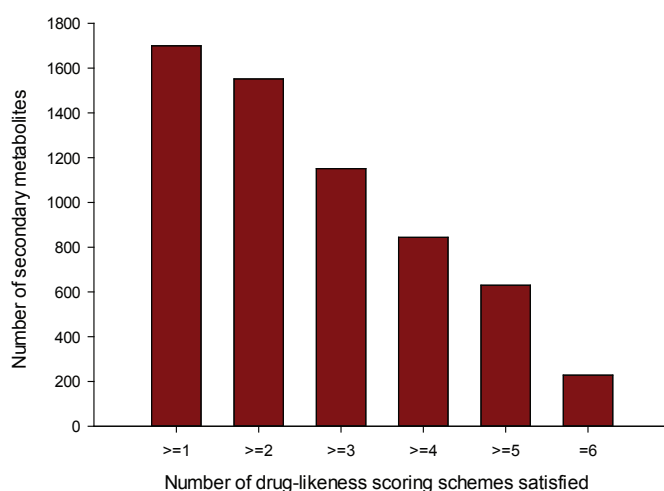


Figure 4.6: Histogram showing the number of secondary metabolites in MeFSAT which satisfy at least 1 (≥ 1), at least 2 (≥ 2), at least 3 (≥ 3), at least 4 (≥ 4), at least 5 (≥ 5) and all 6 ($=6$) of the drug-likeness scoring schemes evaluated here.

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 (Figure 4.7 and Figure 4.8A). Specifically, the CSN of 1830 secondary metabolites has 335 connected components of which 206 are isolated nodes (Figure 4.7). Similarly, the CSN of 228 drug-like secondary metabolites has 94 connected components of which 55 are isolated nodes (Figure 4.8A). 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 also computed the maximum common substructures (MCSs) for the 10 largest connected components within the CSN of 228 drug-like secondary metabolites (Figure 4.8B and Supplementary Table S4.10). 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 chemicals [184]. We

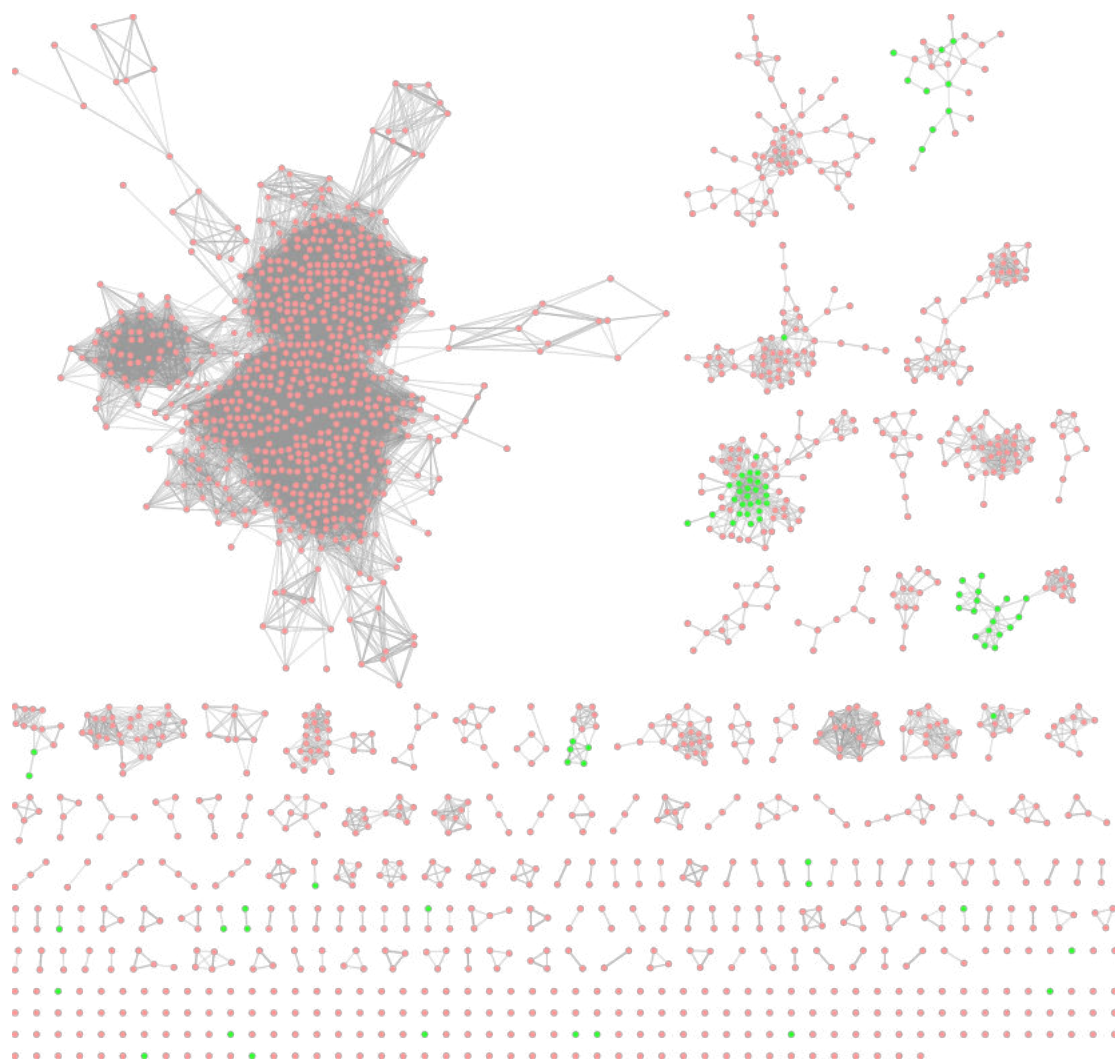


Figure 4.7: Chemical similarity network (CSN) of 1830 secondary metabolites in MeFSAT database. Here, the node color is green if the corresponding secondary metabolite is similar to any of the FDA approved drugs else the node color is pink. Edge thickness is proportional to the computed structural similarity between pairs of secondary metabolites.

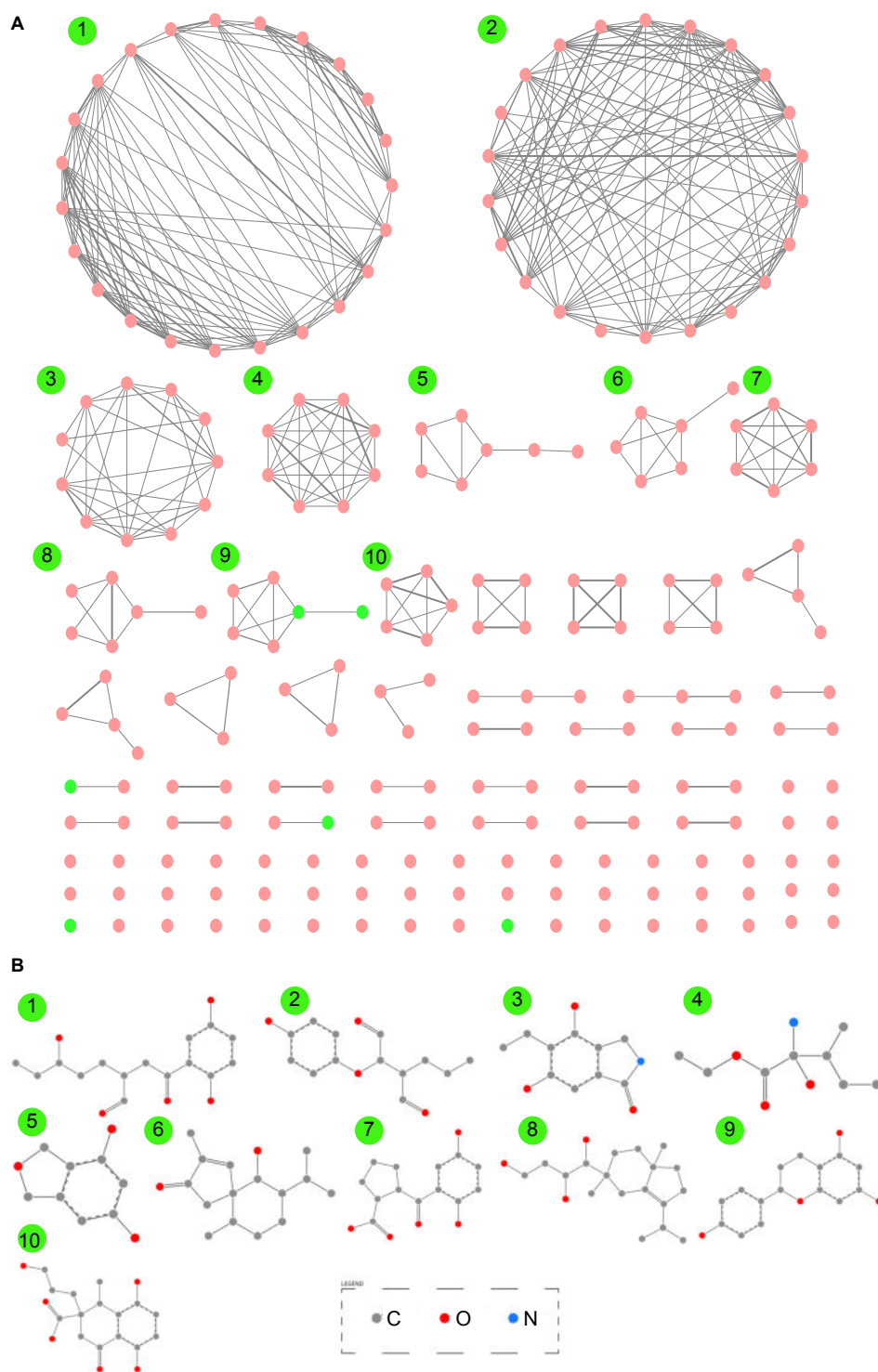


Figure 4.8: (A) Chemical similarity network (CSN) of 228 drug-like secondary metabolites in MeFSAT database. Here, the node color is green if the corresponding secondary metabolite is similar to any of the FDA approved drugs else the node color is pink. Edge thickness is proportional to the computed 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.

used RDKit [74] 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. The computed MCSs have been visualized using SMARTSview [185]. Figure 4.8B displays the unique MCS underlying the 10 largest connected components in the CSN of drug-like secondary metabolites. In future, it will be worthwhile to further explore the scaffold diversity [152] of these drug-like fungal secondary metabolites in MeFSAT.

4.7 Discussion

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 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 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 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.* [84] 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 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 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.

Supplementary Information

Supplementary Tables S4.1-S4.10 associated with this chapter are available for download from the GitHub repository: https://github.com/asamallab/PhDThesis-Vivek_Ananth_RP/blob/main/SI/ST_Chapter4.xlsx.

Chapter 5

***In silico* identification of potential anti-COVID drugs from phytochemicals of Indian medicinal plants**

In December 2019, a new respiratory disease with unknown cause with clinical symptoms of fever, cough, shortness of breath, fatigue and pneumonia was first reported in Wuhan, China [186–188]. While most cases of this new disease showed mild to moderate symptoms, a small fraction of cases, especially those with comorbid conditions like diabetes and hypertension, developed fatal conditions such as acute respiratory distress syndrome (ARDS) due to severe lung damage [189]. In January 2020, a novel beta-coronavirus, initially named 2019-nCoV, was discovered to be the etiological agent of this new disease [186–188]. Subsequently, human-to-human transmission of this disease was confirmed [189]. On 11 February 2020, the international committee on taxonomy of viruses permanently named 2019-nCoV as severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and the WHO named the associated disease as coronavirus disease 2019 (COVID-19). As of 8 June 2022, the number of laboratory confirmed COVID-19 cases and deaths have already surpassed 500 million and 6 million, respectively (<https://covid19.who.int/>). In short, the COVID-19 pandemic has caused an

unprecedented public health and economic emergency for humankind not witnessed in nearly a century.

Coronaviruses are enveloped, positive-sense, single-stranded RNA viruses with large viral genomes. The publication of SARS-CoV-2 genome in January 2020 led to its taxonomic classification into the family *Coronaviridae* and genus *Betacoronavirus* [186–188]. Prior to COVID-19 pandemic, there are two precedences of betacoronaviruses causing outbreaks of severe respiratory disease, namely the severe acute respiratory syndrome coronavirus (SARS-CoV) and the middle east respiratory syndrome coronavirus (MERS-CoV) [190]. In 2002–2003, SARS-CoV emerged in China and led to 8098 infections and 774 deaths across the world [190]. Interestingly, the SARS-CoV-2 genome shares ~80% nucleotide identity with SARS-CoV [186–188]. In 2012, MERS-CoV emerged in Saudi Arabia and led to 2521 infections and 866 deaths that were largely limited to Middle Eastern countries [190]. Unlike SARS and MERS, the geographic spread of COVID-19 and the ensuing mortality is significantly higher. Several small molecules with strong *in vitro* antiviral activity against SARS-CoV-2 have been identified [191, 192] and are under different stages of clinical trials. Remdesivir, one such small molecule, was granted emergency use authorization by US FDA for the treatment of COVID-19 [193]. In present circumstances, an immediate goal of biomedical research is to develop antivirals or anti-COVID therapeutics for SARS-CoV-2 [194].

The SARS-CoV-2 genome comprises ~30,000 bases with 14 open reading frames (ORFs) coding for 27 proteins [195]. The genome organization of SARS-CoV-2 is similar to other betacoronaviruses with the 5'-region coding for non-structural proteins and the 3'-region coding for structural proteins. Important structural proteins of SARS-CoV-2 coded by the 3'-region include the spike (S) surface glycoprotein, the envelope (E) protein, the matrix (M) protein and the nucleocapsid (N) protein [195]. The 5'-region of the SARS-CoV-2 genome contains the replicase gene which codes for two overlapping polyproteins, pp1a and pp1ab, which are proteolytically cleaved by two important non-structural proteins, 3-chymotrypsin like protease (3CL^{pro}) and papain-like protease

(PL^{pro}), to produce functional (non-structural) proteins. Other important non-structural proteins of SARS-CoV-2 for the viral life cycle include the RNA-dependent RNA polymerase (RdRp) and helicase (Nsp13). Accordingly, the 4 non-structural proteins, 3CL^{pro}, PL^{pro}, RdRp and Nsp13, along with the spike glycoprotein of SARS-CoV-2 are among the most attractive targets for anti-COVID drugs [194].

To expedite the search for anti-COVID drugs, several computational studies have used homology modeling or published crystal structures of SARS-CoV-2 proteins and diverse small molecule libraries, to predict potential inhibitors of SARS-CoV-2 proteins including among existing approved drugs for repurposing and natural compounds (see e.g., [196–204]). In comparison, fewer computational studies [196, 199] have focussed on identification of potential inhibitors of host factors. Further, there are reports from China of successful use of traditional Chinese medicine and associated herbs in treatment of COVID-19 patients [205]. On similar lines, there have been suggestions to tap the rich legacy of traditional Indian medicine and information on phytochemicals of Indian herbs in the search for anti-COVID drugs [206].

In this chapter, we present results from our two research studies highlighting the biological application of the curated natural product space of Indian medicinal plants captured in IMPPAT database, for the identification of potential anti-COVID drugs. Specifically, we perform virtual screening of 14011 phytochemicals from Indian medicinal plants to identify potential natural product inhibitors of: (a) key host factors TMPRSS2 and cathepsin L, and (b) SARS-CoV-2 helicase Nsp13. **The work reported in this chapter is contained in the published manuscripts [53, 54].**

5.1 TMPRSS2 and cathepsin L: key human proteases for host cell entry of SARS-CoV-2

Rather than targeting SARS-CoV-2 proteins important for viral life cycle, an alternative approach to anti-COVID drugs involve targeting host factors key to SARS-CoV-2 infec-

tion [194]. For host cell entry, SARS-CoV-2 employs the spike (S) protein whose S1 subunit has a receptor binding domain (RBD) that specifically recognizes the cell surface receptor angiotensin converting enzyme 2 (ACE2) [85,86,207–209]. Hoffmann *et al.* [85] showed that the host cell proteases, Transmembrane Protease Serine 2 (TMPRSS2) and cathepsin L or cathepsin B, can carry out S protein priming required for SARS-CoV-2 entry. In parallel, Ou *et al.* [86] used specific inhibitors of cathepsin L and cathepsin B to show that cathepsin L rather than cathepsin B is essential for S protein priming of SARS-CoV-2 and membrane fusion in lysosomes.

The above studies highlight at least two alternate pathways for host cell entry of SARS-CoV-2. On the one hand, after SARS-CoV-2 attachment to ACE2, the membrane fusion and cytoplasmic entry can occur at the plasma membrane provided the cell surface protease TMPRSS2 is available to carry out S protein priming. On the other hand, after SARS-CoV-2 attachment to ACE2, the virus can be internalized as part of endosomes in the endocytic pathway, and later, the membrane fusion and cytoplasmic entry will occur in lysosomes provided the lysosomal protease cathepsin L is available to carry out S protein priming [85, 86, 207]. Depending on the target cell and associated expression of host cell proteases, SARS-CoV-2 may use one of the alternative pathways for host cell entry. Importantly, the above-mentioned studies also showed that known inhibitors camostat mesylate and nafamostat mesylate of TMPRSS2 and E-64d and PC-0626568 (SID26681509) of cathepsin L can block or significantly reduce the host cell entry of SARS-CoV-2 [85, 86, 207]. In conclusion, human proteases TMPRSS2 and cathepsin L are key factors for host cell entry and are important targets for anti-COVID drugs [85, 86, 207, 210]. Thus, in our first study, we identified potential phytochemical inhibitors of the host cell proteases TMPRSS2 (Figure 5.1A) and cathepsin L (Figure 5.2).

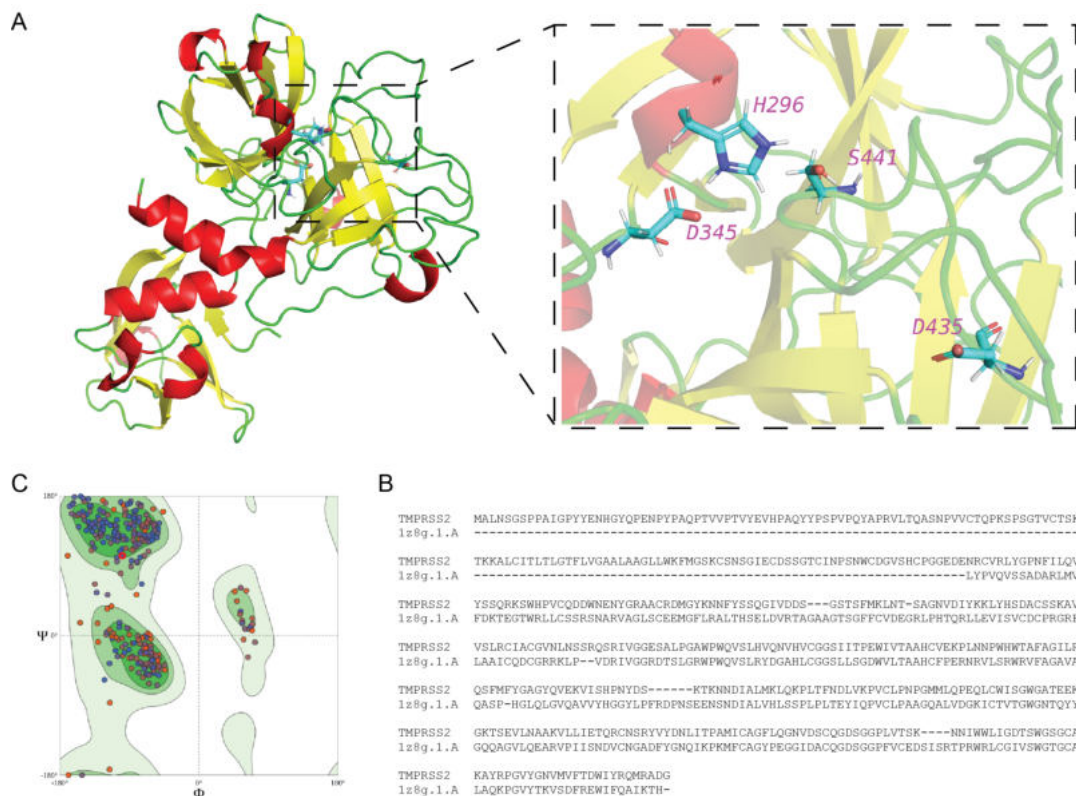


Figure 5.1: (A) Cartoon representation of the homology model structure of TMPRSS2 which has been energy-minimized using UCSF Chimera. The figure zooms into the region containing the catalytic triad Ser441 (S441), His296 (H296) and Asp345 (D345), and the substrate binding residue Asp435 (D435) in the S1 subsite of the enzyme. (B) Alignment of protein sequences for TMPRSS2 and hepsin (PDB 1Z8G) which was used as a template to model the structure of TMPRSS2. (C) General Ramachandran plot of the energy-minimized model structure of TMPRSS2, which displays the torsional angles, phi (ϕ) and psi (ψ), of the amino acid residues in the protein.

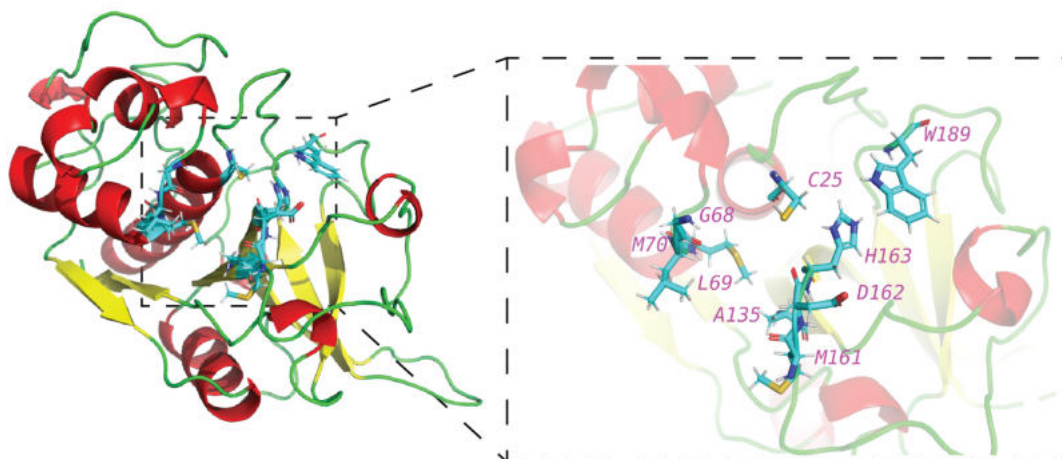


Figure 5.2: Cartoon representation of the crystal structure of human cathepsin L (PDB 5MQY). The figure zooms into the region containing the catalytic residues Cys (C25) and His163 (H163) in the S1 subsite, residues Asp162 (D162), Met161 (M161), Ala135 (A135), Met70 (M70) and Leu (L69) in the S2 subsite, Trp189 (W189) at the centre of S1' subsite and the conserved residue Gly68 (G68) in the S3 subsite of the enzyme.

5.2 SARS-CoV-2 helicase Nsp13

In the second study, we focused on the identification of potential phytochemical inhibitors of the SARS-CoV-2 helicase Nsp13. Nsp13 has both ATPase and helicase activity, as it unwinds the viral RNA helices in an ATP dependent manner [211]. Thus, Nsp13 plays an important role in viral life cycle by enabling correct folding and replication of viral RNA [211]. Notably, due to its high sequence conservation across the coronavirus family, Nsp13 is considered an attractive target for the development of antiviral drugs [87,212]. Also, it was shown that SARS-CoV-2 helicase Nsp13 can hydrolyze all types of NTPs including ATP to unwind the RNA helices [211]. Therefore, the known ATP binding site of the helicase Nsp13 is a promising target for effective inhibition. In this direction, the recently deposited crystal structure of SARS-CoV-2 helicase Nsp13 (PDB 6ZSL) has made development of anti-COVID drugs via targeting of Nsp13 more viable. The Nsp13 of SARS-CoV and SARS-CoV-2 share 99.8% sequence identity [87,213]. Hence, similar to the SARS-CoV Nsp13 structure [212], the SARS-CoV-2 Nsp13 adopts a triangular pyramid shape with five domains namely, the RecA-like domains 1A and 2A, the 2B domain, the zinc-binding domain (ZBD) and the stalk domain (Figure 5.3).

5.3 Methods

In this section, we describe the methods incorporated in the virtual screening workflow for the identification of potential phytochemical inhibitors of: (a) key host factors TMPRSS2 and cathepsin L, and (b) SARS-CoV-2 helicase Nsp13.

5.3.1 Preparation of ligand library of phytochemicals

We compiled a library of 14011 phytochemicals produced by medicinal plants used in traditional Indian medicine, and the main source of this compilation was IMPPAT [50]. Next, the standard drug-likeness measure, Lipinski's rule of five (RO5) [176], was used

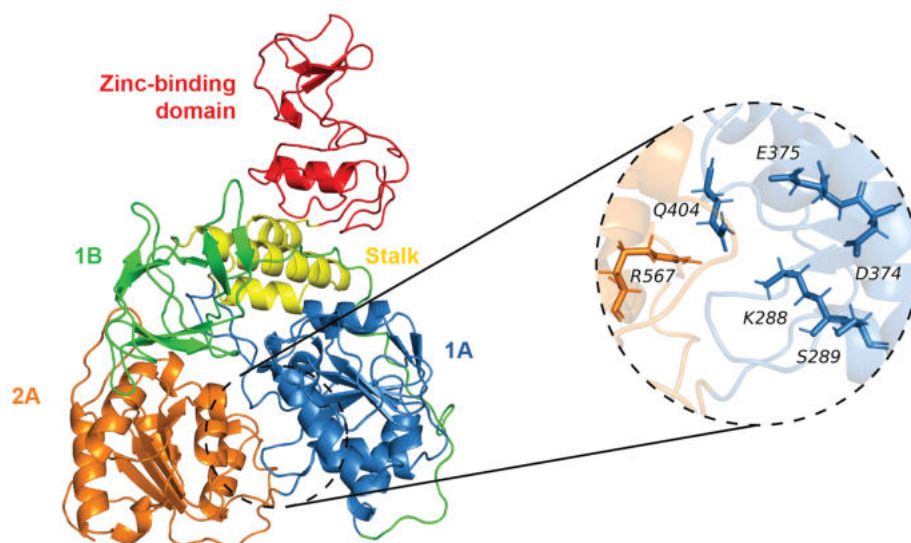


Figure 5.3: Cartoon representation of the prepared crystal structure of SARS-CoV-2 helicase Nsp13 (PDB 6ZSL). The figure shows the five domains in the Nsp13 structure, namely, zinc binding domain (ZBD) colored in red, the stalk domain colored in yellow, the 1B domain colored in green, the RecA-like domains 1A and 2A colored in blue and orange, respectively. The ATP binding site of the Nsp13 with six key residues involved in ATP hydrolysis is shown in expanded view.

to filter a subset of 10510 drug-like phytochemicals. The 3D chemical structures of the filtered 10510 drug-like phytochemicals were obtained from PubChem [214] and subsequently energy minimized using OpenBabel [75].

5.3.2 Molecular docking of the phytochemicals to the target proteins

We performed protein-ligand docking of the phytochemicals to the target proteins using AutoDock Vina [215]. The 3D structures of prepared ligands in .pdb format were converted to .pdbqt format using the python script prepare_ligand4.py [216]. Similarly, the protein structures of target proteins namely, TMPRSS2, cathepsin L and Nsp13, in .pdb format were converted to .pdbqt format using the python script prepare_receptor4.py [216]. For protein-ligand docking, an appropriate grid box was manually placed considering the key residues in target proteins such as the catalytic residues and substrate binding residues as reported in the literature. For all the target proteins, the molecular docking of prepared ligands was performed using AutoDock Vina with the exhaustiveness set as

8. The top conformation of the docked ligand with lowest binding energy, i.e., the best docked pose, was obtained from the output of AutoDock Vina using the associated script `vina_split`. Subsequently, a combined structure file in `.pdb` format of the docked protein-ligand complex (with ligand in the best docked pose) was prepared using a custom script and `pdb-tools` [217].

5.3.3 Identification of protein-ligand interactions

The combined structure of docked protein-ligand complex in `.pdb` format was used to determine the ligand binding site residues in the target protein and different non-covalent interactions. These non-covalent protein-ligand interactions were identified using different geometric criteria which are specific to different types of interactions.

Binding site residue. Ligand binding site residues are defined as amino acids in protein which have at least one non-hydrogen atom in the proximity of at least one non-hydrogen atom of the ligand. The distance cut off to determine this proximity between non-hydrogen atoms of protein and ligand is taken to be the sum of their van der Waals radius plus 0.5 Å [218].

Hydrogen bonds. The accepted geometric criteria for hydrogen bonds of type D-H...A are as follows. Firstly, the distance between the hydrogen (H) and acceptor (A) atom should be less than the sum of their van der Waals radii. Secondly, the angle formed by donor (D), H and A atoms should be $> 90^\circ$ (Figure 5.4A). Moreover, carbon (C), nitrogen (N), oxygen (O) or sulfur (S) atoms can be donors while N, O or S atoms can be acceptors [219–221].

Chalcogen bonds. In contrast to hydrogen bonds, chalcogen bonds are of type C-Y...A, where Y can be a S or selenium (Se) atom and A can be a N, O or S atom. The accepted geometric criteria for chalcogen interactions are as follows. Firstly, the distance between Y and A should be less than the sum of their van der Waals radii. Secondly, the angle formed by the triad, that is $\angle C-Y \cdots A$, should lie in the range 150° to 180° (Figure

5.4B) [222].

Halogen bonds. Halogen bonds are of type C-Y...A-B, where halogen Y can be a Fluorine (F), Chlorine (Cl), Bromine (Br) or Iodine (I) atom and A can be a N, O or S atom. The formation of the halogen bond is favoured when the distance between Y and A is $\leq 3.7 \text{ \AA}$ and the angle θ_1 of the A atom relative to the C-Y bond, and the angle θ_2 of the halogen Y relative to the A-B bond should be $\geq 90^\circ$ (Figure 5.4C) [223].

π - π stacking. This interaction occurs between two aromatic rings and can be majorly classified into two types, namely, face-to-face and face-to-edge. In the case of face-to-face type of π - π interaction, the distance between the centroids of the two participating aromatic rings should be $< 4.4 \text{ \AA}$ and the angle between their ring planes should be $< 30^\circ$. In the case of face-to-edge type of π - π interaction, the distance between the centroids of the two participating aromatic rings should be $< 5.5 \text{ \AA}$ and the angle formed by the ring planes should be in the range 60° to 120° (Figure 5.4D,E).

Hydrophobic interactions. The geometric criteria for the formation of hydrophobic interactions between atoms in protein and ligand are as follows [224]. The distance between a carbon atom in protein or ligand and a carbon, halogen or sulfur atom in ligand or protein, respectively, should be $\leq 4 \text{ \AA}$. Furthermore, we ensure that the involved atoms in a hydrophobic interaction between protein and ligand do not form hydrogen, chalcogen or halogen bonds between them [224].

In order to detect the above-mentioned protein-ligand interactions, an in-house Python program was written to enable batch processing of combined structure files containing docked protein-ligand complexes for our large phytochemical library.

5.3.4 Molecular Dynamics simulations

We performed molecular dynamics (MD) simulations of the protein and the protein-ligand complex using GROMACS 5.1.5 [225] with GROMOS96 54a7 force field [226]. The Automated Topology Builder (ATB) version 3.0 (<https://atb.uq.edu.au/>) was used to

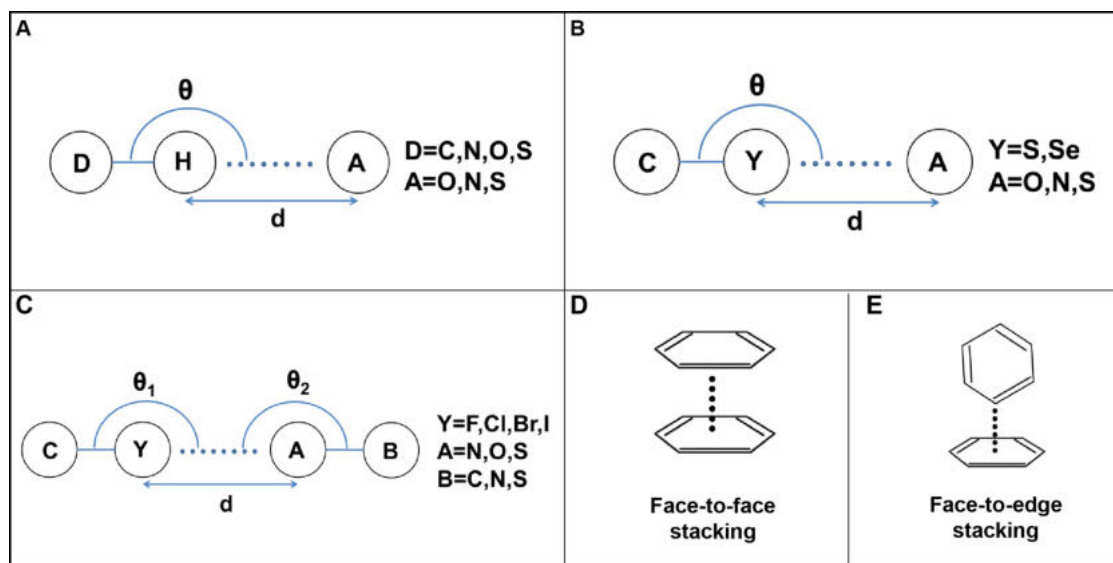


Figure 5.4: Geometric criteria for the identification of protein-ligand interactions. (A) Hydrogen bond, (B) Chalcogen bond, (C) Halogen bond, (D) face-to-face $\pi - \pi$ stacking, and (E) face-to-edge $\pi - \pi$ stacking.

generate topology for the top inhibitors [227]. To prepare the system for MD simulation, the uncomplexed protein or the protein–ligand complex was placed at the center of a cubic box with periodic boundary conditions, and thereafter, the system was solvated by adding simple point-charge (SPC) water and neutralized by Na^+ and Cl^- ions. Then the system was heated to 310 K during a constant number of particles, volume, and temperature (NVT) simulation. Then the pressure was equilibrated to 1 bar during a constant number of particles, pressure, and temperature (NPT) simulation. In both the above simulations, protein and ligand were position restrained. Thereafter, a production MD run was performed, after removing the position restraint on protein and ligand. During the MD simulation, the structural coordinates were saved after every 2 ps. Further, the system temperature and pressure were maintained at 310 K and 1 bar using v-rescale temperature [228] and Parrinello-Rahman pressure coupling method [229], respectively. For the temperature and pressure coupling, the time constant was kept at 0.1 ps and 2 ps, respectively. The atom pairs within the cut off of 1.4 nm were used for the computation of short-range interactions. Whereas, Particle mesh Ewald (PME) method with fourth order cubic interpolation and 0.16 nm grid spacing were used for the computation of long-range

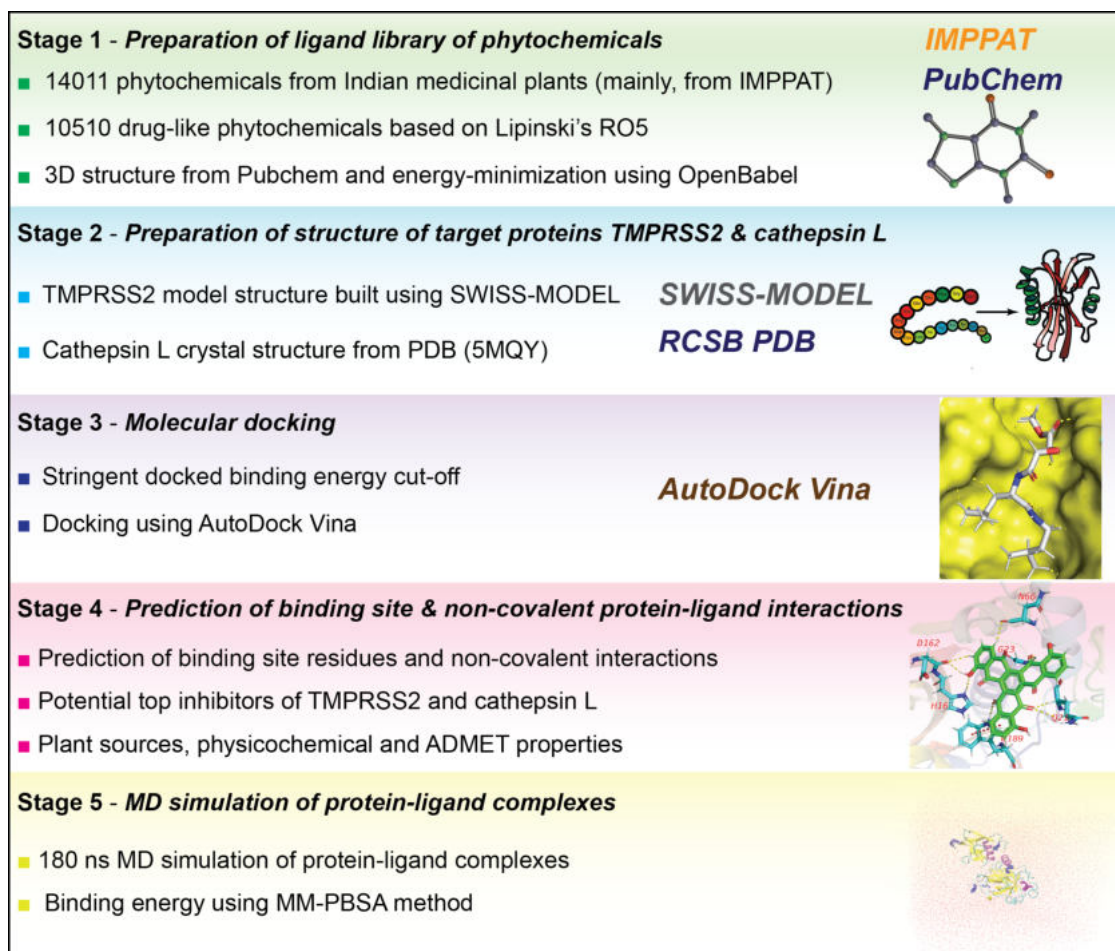


Figure 5.5: Workflow to identify potential phytochemical inhibitors of human proteases TM-PRSS2 and cathepsin L.

electrostatic interactions. All bonds were constrained using the LINCS method. The trajectories obtained from the MD simulations were then used to compute and analyze the radius of gyration of the protein (R_g), root mean square deviation (RMSD) of the protein backbone C_α atoms and root mean square fluctuation (RMSF) of the protein backbone C_α atoms using GROMACS scripts.

5.4 Virtual screening for key host factors in SARS-CoV-2 infection

In this section we describe the identification of potential phytochemical inhibitors of the host factors TM-PRSS2 and cathepsin L (Figure 5.5).

In the first stage, we prepared the 10510 drug-like phytochemicals from Indian medicinal plants as described in Section 5.3.1. In the second stage, we prepared the host proteins TMPRSS2 and cathepsin L for docking with the prepared ligands. TMPRSS2 is a trypsin-like serine protease whose catalytic site consists of the triad Ser441 (S441), His296 (H296) and Asp345 (D345) [230]. It is well established that trypsin-like serine proteases cleave peptide bonds following positively charged amino acid residues such as arginine or lysine, and this specificity of the enzyme is determined by a negatively charged aspartate residue located at the bottom of its S1 pocket [231]. In TMPRSS2, this specificity is determined by the conserved negatively charged residue Asp435 (D435) at the bottom of the S1 pocket [230].

At the time of performing this study, the experimentally determined 3D structure of TMPRSS2 was not available (as of 01 July 2020), and thus, we used SWISS-MODEL [232] webserver (<https://swissmodel.expasy.org/interactive>) to perform homology modeling of TMPRSS2 using crystal structure of human protein hepsin (PDB 1Z8G) [233] as the template (Figure 5.1A). Note that hepsin (PDB 1Z8G) is also a Type II transmembrane trypsin-like serine protease, and it shares 38% sequence similarity with TMPRSS2 (NP_005647.3) (Figure 5.1B). Subsequently, UCSF Chimera [234] was used to minimize the energy of the TMPRSS2 model structure obtained from SWISS-MODEL. In the TMPRSS2 model, 94.19% of the amino acid residues were found to be in the Ramachandran favoured regions in the Ramachandran plot (Figure 5.1C) and the model structure has a MolProbity [235] score of 1.50. Figure 5.1 displays the prepared TMPRSS2 model structure with the catalytic triad S441, H296 and D345 and the substrate binding residue D435 in the S1 subsite.

For cathepsin L, the crystal structure (PDB 5MQY) with 1.13 Å resolution was used after energy-minimization using UCSF Chimera (Figure 5.2). Figure 5.2 shows the prepared cathepsin L structure with important residues in S1, S2 and S1' subsites of the enzyme [236]. Previous research has revealed that S1 and S2 subsites of cathepsin L are important for the specificity of the enzyme [236, 237]. In cathepsin L, the catalytic

site consists of Cys25 (C25) and His163 (H163) in the S1 subsite, and Trp189 (W189) is at the center of the S1' subsite [236] (Figure 5.2). In cathepsin L, the S2 subsite with important residues Asp162 (D162), Met161 (M161), Ala135 (A135), Met70 (M70) and Leu69 (L69) forms a deep hydrophobic pocket, and lastly, the conserved residue Gly68 (G68) is at the center of the S3 subsite [236,238] (Figure 5.2).

In the third stage, we performed protein-ligand docking using AutoDock Vina as described in Section 5.3.2. To decide on a stringent binding energy cut off for the identification of potential inhibitors, docking was first performed for known inhibitors of target proteins. Based on the docking binding energies of the known inhibitors, camostat and nafamostat, to TMPRSS2, we decided on a stringent criteria of binding energy ≤ -8.5 kcal/mol for the best docked pose of screened ligands to identify potential inhibitors of TMPRSS2. Similarly, based on the docking binding energies of the known inhibitors, E-64d and PC-0626568, to cathepsin L, we decided on a stringent criteria of binding energy ≤ -8.0 kcal/mol for the best docked pose of screened ligands to identify potential inhibitors of cathepsin L. The description on the docking of the known inhibitors to TMPRSS2 and cathepsin L to decide on the binding energy cut off is provided in Appendix A.

Thereafter, molecular docking was performed for the prepared ligands in the phytochemical library against the prepared structures of TMPRSS2 and cathepsin L (Section 5.3.2). We then filtered the subset of phytochemicals whose binding energy in the best docked pose with TMPRSS2 (respectively, cathepsin L) is ≤ -8.5 kcal/mol (respectively, ≤ -8.0 kcal/mol). Moreover, the best docked pose with TMPRSS2 or cathepsin L of each filtered phytochemical was separated from AutoDock Vina output file, and then, combined with the target protein structure to obtain the docked protein-ligand complex (Section 5.3.2). At the end of third stage, we obtained 101 phytochemicals whose binding energy in the best docked pose with TMPRSS2 is ≤ -8.5 kcal/mol and 16 phytochemicals whose binding energy in the best docked pose with cathepsin L is ≤ -8.0 kcal/mol.

In the fourth stage, the structure of docked protein-ligand complex for each filtered

phytochemical from third stage was used to determine the ligand binding site residues in the target protein and different non-covalent interactions such as hydrogen bond, halogen bond, hydrophobic interactions, etc. between ligand and target protein (Section 5.3.3). In case of TMPRSS2, the specificity of this trypsin-like protease is determined by the conserved substrate binding residue D435 in the S1 pocket [230], and therefore, a potent inhibitor should either bind to or form non-covalent interactions with D435. In case of cathepsin L, the specificity of this cysteine protease is dependent on the catalytic residues, C25 and H163, in the S1 subsite, and therefore, a potent inhibitor should either bind to or form non-covalent interactions with the catalytic residues. In this work, we consider a phytochemical to be a potential inhibitor of TMPRSS2 only if the ligand binding energy in the best docked pose is ≤ -8.5 kcal/mol and the ligand binds to or forms non-covalent interactions with the residue D435 in TMPRSS2. Similarly, we consider a phytochemical to be a potential inhibitor of cathepsin L only if the ligand binding energy in the best docked pose is ≤ -8.0 kcal/mol and the ligand binds to or forms non-covalent interactions with the residues C25 and H163 in cathepsin L.

At the end of fourth stage, we obtained 96 phytochemicals (labelled T1–T96; Figure 5.6; Supplementary Table S5.1) as potential inhibitors of TMPRSS2 and 9 phytochemicals (labelled C1–C9; Figure 5.8; Supplementary Table S5.2) as potential inhibitors of cathepsin L. Using IMPPAT database [50], we provide a list of Indian medicinal plants that can produce the identified phytochemical inhibitors of TMPRSS2 and cathepsin L. Furthermore, we have also compiled information on potential antiviral or anti-inflammatory use in traditional medicine of the herbal sources of the identified phytochemical inhibitors of TMPRSS2 and cathepsin L (Table 5.1 and Table 5.2; Supplementary Table S5.3). In Supplementary Tables S5.4 and S5.5 we list the ligand binding site residues and non-covalent protein-ligand interactions for the identified phytochemical inhibitors of TMPRSS2 and cathepsin L. We also provide the physicochemical and predicted ADMET properties [135,239] of the identified phytochemical inhibitors of TMPRSS2 and cathepsin L (Supplementary Tables S5.6 and S5.7).

Finally, in the fifth stage, for the top three inhibitors of TMPRSS2 namely, T1 (qingdainone), T2 (edgeworoside C) and T3 (adlumidine), and of cathepsin L namely, C1 (ararobinol), C2 ((+)-oxoturkiyenine) and C3 (3 α ,17 α -cinchophylline), their respective protein-ligand complexes were analyzed using MD simulation and their interaction binding energy was computed using MM-PBSA method. The results from the MD simulation and MM-PBSA method based binding energy computations are described in Appendix A.

5.4.1 Potential Phytochemical Inhibitors of TMPRSS2

As mentioned above, we identified 96 potential natural product inhibitors of TMPRSS2 by computational screening of 14011 phytochemicals produced by Indian medicinal plants, and these 96 compounds labelled T1-T96 are listed in Supplementary Table S5.1 along with their PubChem identifier, chemical name, IUPAC name and structure in SMILES format. In this section, we provide a detailed discussion of the top nine phytochemical inhibitors (labelled as T1–T9) whose binding energies in the best docked poses with TMPRSS2 are ≤ -9.2 kcal/mol. Figure 5.6 displays the structure of these top 9 phytochemical inhibitors of TMPRSS2 and Table 5.1 provides a list of Indian medicinal plants that can produce them.

Phytochemical T1, qingdainone, has a binding energy of -9.6 kcal/mol. T1 is a quina-zoline alkaloid produced by *Strobilanthes cusia*, a herb with antiviral activity [240]. Figure 5.7A shows the TMPRSS2 residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with T1. TMPRSS2 residue D440 forms C-H \cdots O type hydrogen bond with T1 whereas residue A399 forms C-H \cdots N type hydrogen bond with T1. Further, T1 forms hydrophobic contacts with residues I381, S382, T387, E388, N398, A400, D440, C465 and A466.

Phytochemical T2, edgeworoside C, also has a binding energy of -9.6 kcal/mol. T2 is a coumarin produced by *Edgeworthia gardneri*, a medicinal plant consumed as a herbal

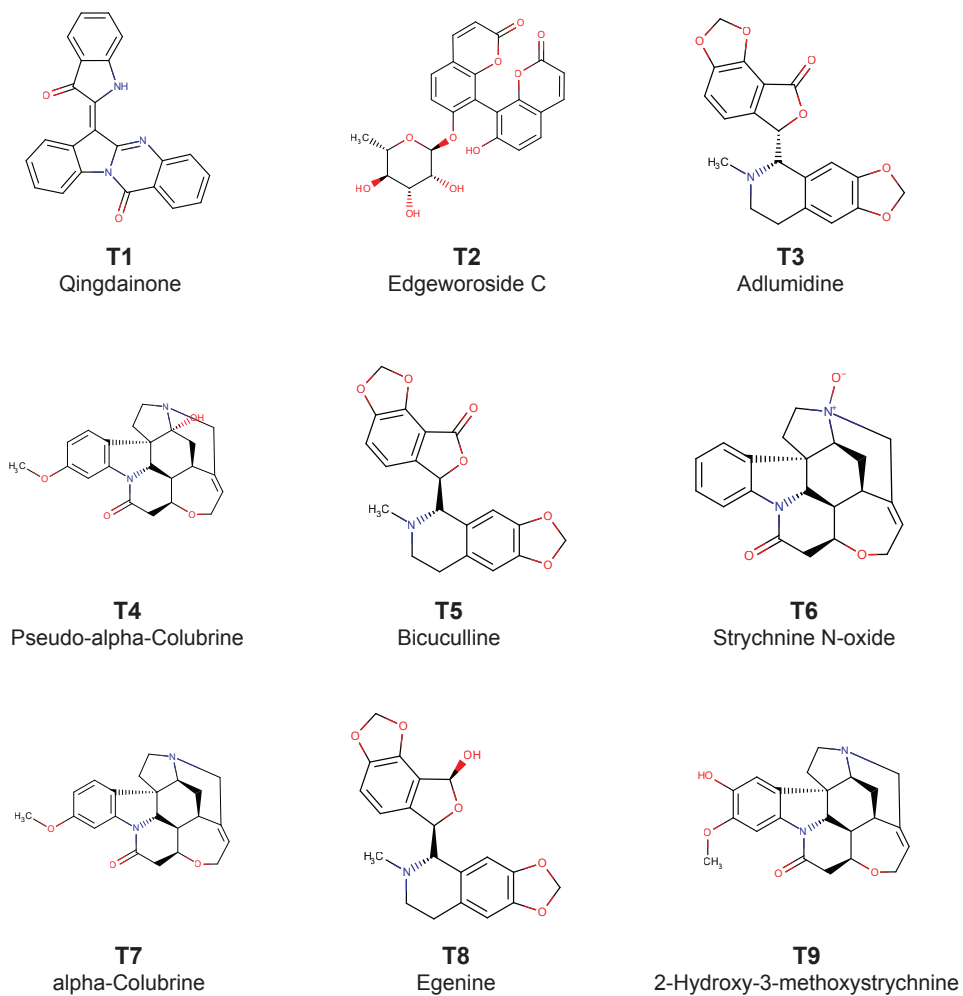
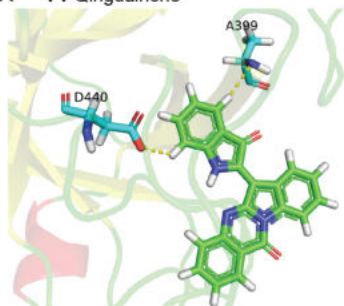
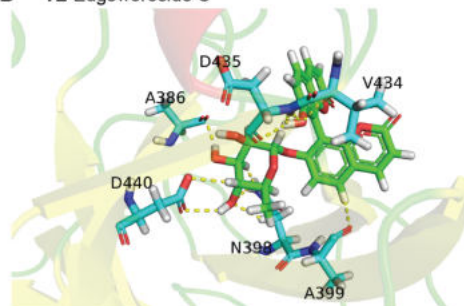


Figure 5.6: Molecular structure and chemical name of the top 9 phytochemical inhibitors (compounds T1–T9) of Tmprss2.

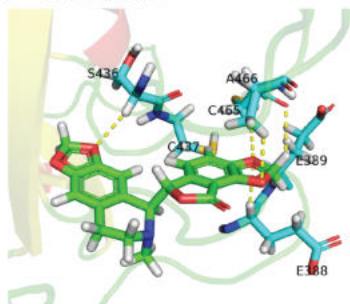
A T1 Qingdaine



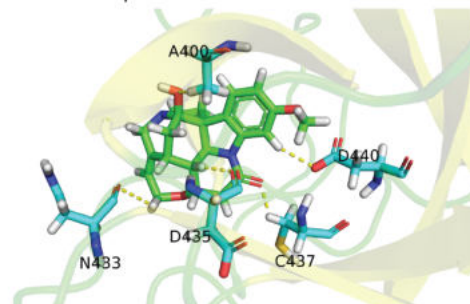
B T2 Edgeworoside C



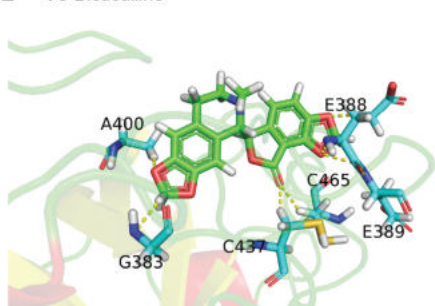
C T3 Adlumidine



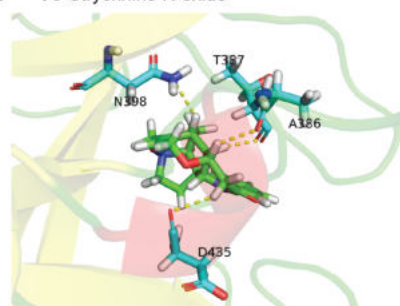
D T4 Pseudo-alpha-Colubrine



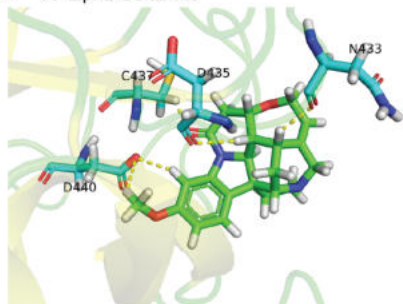
E T5 Bicuculline



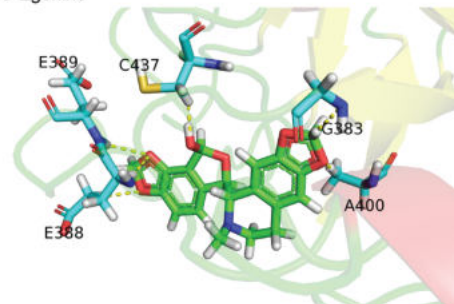
F T6 Strychnine N-oxide



G T7 alpha-Colubrine



H T8 Egenine



I T9 2-Hydroxy-3-methoxystrychnine

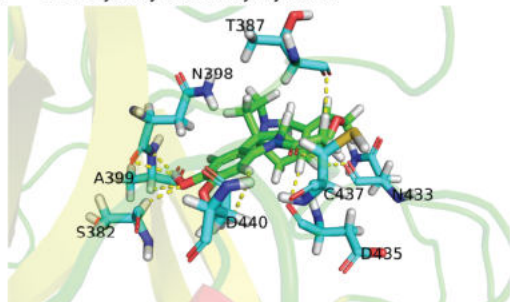


Figure 5.7 (previous page): Cartoon representation of the protein-ligand interactions of the phytochemical inhibitors of TMPRSS2. Interactions of TMPRSS2 residues with atoms of (A) T1, (B) T2, (C) T3, (D) T4, (E) T5, (F) T6, (G) T7, (H) T8 and (I) T9. The carbon atoms of the ligand are shown in green colour while the carbon atoms of the amino acid residues in TMPRSS2 are shown in cyan colour. TMPRSS2 residues interacting with the ligand atoms via hydrogen bonds or $\pi - \pi$ stacking are labelled with their corresponding one letter amino acid code along with their residue number in the protein sequence. The hydrogen bonds and $\pi - \pi$ stacking are displayed using yellow and red dotted lines, respectively.

tea in Tibet [241]. In traditional medicine, *Edgeworthia gardneri* has been used to treat metabolic disorders including diabetes [242, 243]. Figure 5.7B shows the TMPRSS2 residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with T2. T2 forms 12 hydrogen bonds with residues A386, N398, A399, V434, D435 and D440 of TMPRSS2. The phenolic hydroxyl group of T2 acts as both acceptor and donor forming O-H...O and N-H...O type hydrogen bonds with the substrate binding residue D435 and C-H...O type hydrogen bond with residue V434. The hydroxyl groups attached to the pyran ring of T2 form hydrogen bonds with residues A386, N398 and D440. Further, T2 forms hydrophobic contacts with residues E260, I381, A400, N433, and A466.

Phytochemical T3, adlumidine, also has a binding energy of -9.6 kcal/mol. T3 is produced by *Fumaria indica*, a herb used in traditional medicine to treat fever, cough, skin ailments and urinary diseases [244]. Adlumidine has also been reported to be an inhibitor of GABA receptor [245]. Figure 5.7C shows the TMPRSS2 residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with T3. The two 1,3-dioxole groups present in T3 facilitate the formation of an extensive hydrogen bond network with E388, E389, S436, C465 and A466. T3 also forms C-H...S type hydrogen bond [219] with C437. Further, T3 forms hydrophobic contacts with residues E260, I381, S382, T387, N398, A399 and A400.

Phytochemicals T4 (pseudo- α -colubrine), T6 (strychnine N-oxide), T7 (α -colubrine) and T9 (2-hydroxy-3-methoxystrychnine) have binding energies of -9.3 kcal/mol, -9.3 kcal/mol, -9.2 kcal/mol and -9.2 kcal/mol, respectively. These four phytochemicals are monoterpenoid indole alkaloids produced by *Strychnos nux-vomica*. The herb *Strychnos*

nux-vomica is used in traditional Indian medicine and its alkaloids have been shown to exhibit anti-inflammatory, anti-oxidant, antitumor and hepatoprotective activities [246]. Note that *Strychnos nux-vomica* is a poisonous plant whose seeds are extensively used in Ayurveda only after proper detoxification procedure called Shodhana described in Ayurvedic texts [246]. Figure 5.7D shows the TMPRSS2 residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with T4. T4 forms C-H \cdots O type hydrogen bonds with residues A400, N433, D435 (substrate binding residue), C437 and D440. Further, T4 forms hydrophobic contacts with residues E260, I381, S382, T387, N398, A399, V434, D440 and A466. Figure 5.7F shows the TMPRSS2 residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with T6. T6 forms a C-H \cdots N type hydrogen bond with residue N398. The substrate binding residue D435 also forms a C-H \cdots O type hydrogen bond with T6. Further, T6 forms hydrophobic contacts with residues N398, A400, V434 and A466. Figure 5.7G shows the TMPRSS2 residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with T7. T7 forms five C-H \cdots O type hydrogen bonds with residues N433, D435 (substrate binding residue), C437 and D440. Further, T7 forms hydrophobic contacts with residues E260, T387, N398, A399, A400, V434 and A466. Figure 5.7I shows the TMPRSS2 residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with T9. The phenolic hydroxyl group of T9 forms hydrogen bonds with residues S382 and A399. The substrate binding site D435 also forms a C-H \cdots O type hydrogen bond with T9. Further, T9 forms hydrophobic contacts with residues E260, N398, A399, A400 and V434.

Phytochemical T5, bicuculline, has a binding energy of -9.3 kcal/mol, and it is a stereoisomer of T3. T5 is an isoquinoline alkaloid and is produced by herbs *Corydalis govaniiana*, *Nerium oleander* and *Fumaria indica*. Bicuculline has also been reported to be a GABA receptor inhibitor [247]. Figure 5.7E shows the TMPRSS2 residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with T5. The two 1,3-dioxole groups present in T5 facilitate the formation of an extensive hydrogen bond network with residues E388, E389 and A400. The Furan-2-one ring also forms a C-H \cdots O type hydrogen bond with

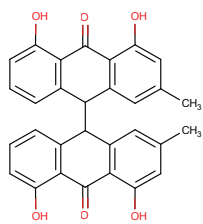
C437. Further, T5 forms hydrophobic contacts with residues E260, T387, E388, N398, A399 and A466.

Phytochemical T8, egenine, has a binding energy of -9.2 kcal/mol. T8 is an isoquinoline alkaloid produced by *Fumaria vaillantii*. In traditional medicine, *Fumaria vaillantii* has been reported to exhibit antifungal, anti-inflammatory and anti-psychotic activities [93]. Figure 5.7H shows the TMPRSS2 residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with T8. The two 1,3-dioxole groups present in T8 form hydrogen bonds with G383, E388, E389 and A400. One of the hydroxyl groups present in T8 forms C-H \cdots O type hydrogen bond with residue C437. Further, T8 forms hydrophobic contacts with residues T387, A399, E388, N398, E260, and A466.

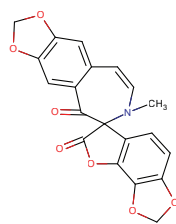
5.4.2 Potential Phytochemical Inhibitors of Cathepsin L

As described we identified nine potential natural product inhibitors of cathepsin L by computational screening of 14011 phytochemicals produced by Indian medicinal plants, and these compounds labelled C1–C9 are listed in Supplementary Table S5.2 along with their PubChem identifier, chemical name, IUPAC name and structure in SMILES format. Figure 5.8 displays the structure of these top nine phytochemical inhibitors of cathepsin L and Table 5.2 provides a list of Indian medicinal plants that produce them.

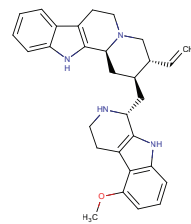
Phytochemical C1, ararobinol, has a binding energy of -8.9 kcal/mol. C1 is a bianthraquinone produced by herb *Senna occidentalis* used in Ayurveda. In traditional medicine, *Senna occidentalis* has been reported for antibacterial, antifungal, anti-inflammatory, anti-diabetic and anti-cancer activities [248]. Interestingly, there is a Chinese patent application [249] on potential use of ararobinol to treat human influenza virus infections, however, this suggests only a potential antiviral activity of C1 not specific to SARS-CoV-2 which further needs to be verified through *in vitro* and *in vivo* experimental studies. Figure 5.9A shows the cathepsin L residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with C1. Residues Q19 and A138 form C-H \cdots N and C-H \cdots O



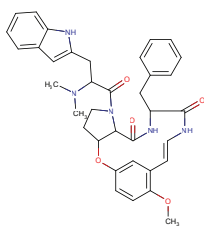
C1
Ararobinol



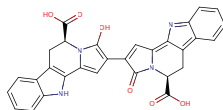
C2
(+)-Oxoturkiyene



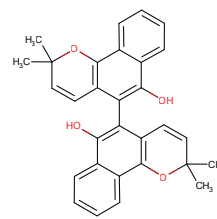
C3
3Alpha,17Alpha-Cinchophylline



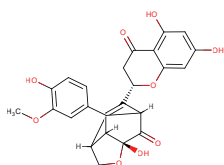
C4
Rugosanine B



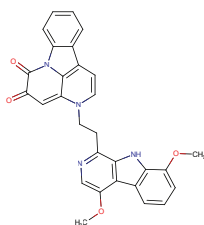
C5
Trichotomine



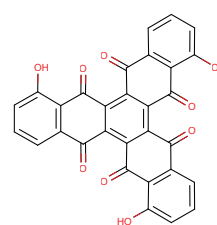
C6
Tectol



C7
Silymonin



C8
Picrasidine M



C9
Trisjuglone

Figure 5.8: Molecular structure and chemical name of the top 9 phytochemical inhibitors (compounds C1–C9) of cathepsin L.

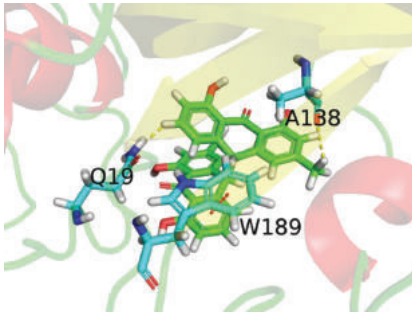
type of hydrogen bonds, respectively, with C1. Also, the residue W189 forms both face-to-edge and face-to-face type of $\pi - \pi$ stacking interaction with C1. Further, C1 forms hydrophobic contacts with residues C25, G139, L144, H163 and W189.

Phytochemical C2, (+)-oxoturkiyenine, has a binding energy of -8.3 kcal/mol. C2 is an isoquinoline-derived alkaloid produced by the herb *Hypocoum pendulum* [250]. Figure 5.9B shows the cathepsin L residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with C2. The 2,5-dihydro-furan ring present in C2 forms two N-H \cdots O type hydrogen bonds with residue Q19 and W189. The catalytic residue H163 forms N-H \cdots O type hydrogen bond with C2. Also, the residue W189 forms a face-to-edge type of $\pi - \pi$ stacking interaction with C2. Further, C2 forms hydrophobic contacts with residues G139, H140, H163 and W189.

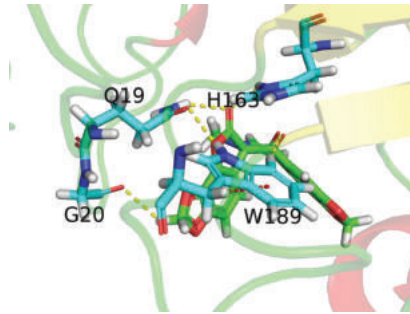
Phytochemical C3, 3 α ,17 α -cinchophylline, has a binding energy of -8.3 kcal/mol. C3 is a cinchophylline-type of alkaloid produced by the herb *Cinchona calisaya*. The Cinchona alkaloids have been reported for their antimicrobial, antiparasitic and anti-inflammatory activities [251]. Figure 5.9C shows the cathepsin L residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with C3. C3 forms 8 hydrogen bonds with residues of cathepsin L. One of the catalytic residue C25 forms C-H \cdots S and N-H \cdots S type hydrogen bonds with C3. The other catalytic residue H163 forms C-H \cdots N and N-H \cdots N type hydrogen bonds with C3. Further, one of the two pyrrole rings present in C3 forms hydrogen bond with residue G23. Lastly, M70 forms a C-H \cdots S type hydrogen bond with C3. Further, C3 forms hydrophobic contacts with residues Q21, C22, L69, M70, A135 and W189.

Phytochemical C4, rugosanine B, has a binding energy of -8.2 kcal/mol. C4 is a cyclopeptide alkaloid produced by the bark of *Ziziphus rugosa* [252]. Various parts of *Ziziphus rugosa* have been reported for their antibacterial, antifungal, anti-inflammatory and analgesic activities [253]. Figure 5.9D shows the cathepsin L residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with C4. The pyrrole ring present in C4 forms a N-H \cdots O type hydrogen bond with residue D162. Moreover, C4 forms C-H \cdots O

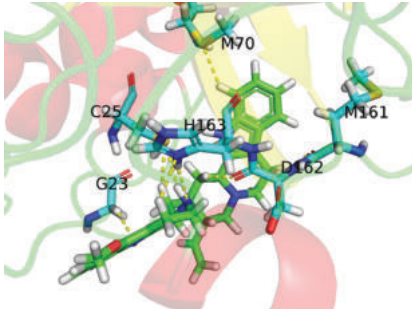
A C1 Ararobinol



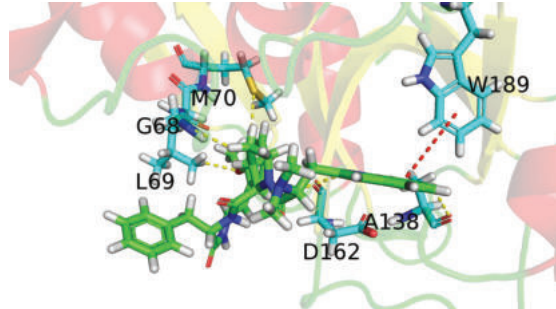
B C2 (+)-Oxoturkiyenine



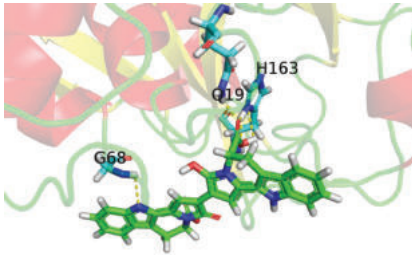
C C3 3Alpha,17Alpha-Cinchophylline



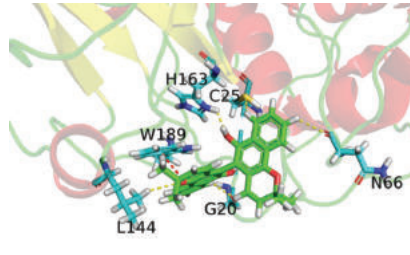
D C4 Rugosanine B



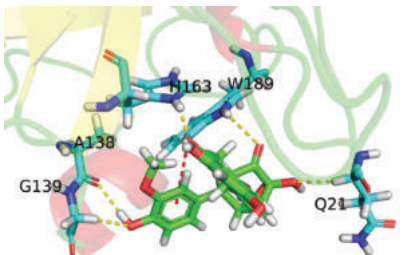
E C5 Trichotomine



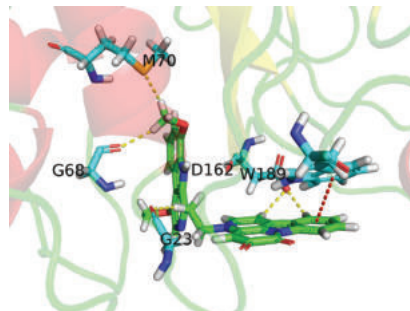
F C6 Tectol



G C7 Silymonin



H C8 Picrasidine M



I C9 Trisjuglone

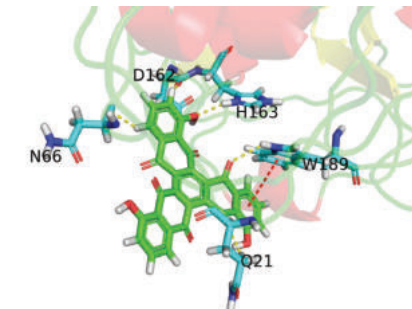


Figure 5.9 (previous page): Cartoon representation of the protein-ligand interactions of the phytochemical inhibitors of cathepsin L. Interactions of cathepsin L residues with atoms of (A) C1, (B) C2, (C) C3, (D) C4, (E) C5, (F) C6, (G) C7, (H) C8 and (I) C9. The carbon atoms of the ligand are shown in green colour while the carbon atoms of the amino acid residues in cathepsin L are shown in cyan colour. Cathepsin L residues interacting with the ligand atoms via hydrogen bonds or $\pi - \pi$ stacking are labelled with their corresponding one letter amino acid code along with their residue number in the protein sequence. The hydrogen bonds and $\pi - \pi$ stacking are displayed using yellow and red dotted lines, respectively.

type hydrogen bonds with A138, D162 and L69. Also, the residue W189 forms a face-to-edge type of $\pi - \pi$ stacking interaction with C4. Further, C4 forms hydrophobic contacts with residues G23, C25, G67, M70, A135, A138, D162, H163, G164, W189 and A214.

Phytochemical C5, trichotomine, has a binding energy of -8.2 kcal/mol. C5 is a bisindole alkaloid present in *Clerodendrum trichotomum*. *Clerodendrum trichotomum* has been reported for its use to treat cold, hypertension, rheumatism, dysentery and other inflammatory diseases [254]. Figure 5.9E shows the cathepsin L residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with C5. The carboxylic acid group present in C5 forms hydrogen bonds with residues Q19 and H163. The indole ring of C5 forms a N-H \cdots N type hydrogen bond with residue G68. Further, C5 forms hydrophobic contacts with residues G23, C25, G67, G68, L69 and Y72.

Phytochemical C6, tectol, has binding energy of -8.1 kcal/mol. C6 is a naphthoquinone derivative [255] present in *Tectona grandis* and *Tecomella undulata*. *Tectona grandis* has been reported to have anti-inflammatory and antiparasitic activities [93]. *Tecomella undulata* has been used to treat syphilis and also reported to have anti-inflammatory and anti-HIV activities [256]. Additionally, Tectol has been reported to inhibit farnesyltransferase enzyme [257]. Figure 5.9F shows the cathepsin L residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with C6. The pyran group of C6 is involved in a C-H \cdots O type hydrogen bond with residue L144. The catalytic residue C25 forms a C-H \cdots S type hydrogen bond with C6. The other catalytic residue H163 forms a N-H \cdots O type hydrogen bond with C6. Also, the residue W189 forms both face-to-face and face-to-edge type of $\pi - \pi$ stacking interaction with C6. Further, C6 forms

hydrophobic contacts with G23, L144 and W189.

Phytochemical C7, silymonin, has a binding energy of -8.1 kcal/mol. C7 is a flavanolignan [258] present in *Silybum marianum*. *Silybum marianum* has been used as a hepatoprotective agent and is reported to have anti-oxidant and anti-inflammatory activities [259]. Figure 5.9G shows the cathepsin L residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with C7. C7 has four hydroxyl groups which help in the formation of an extensive network of hydrogen bonds with residues Q21, A138 and G139. C7 also forms two N-H \cdots O type hydrogen bonds with H163 and W189. Also, the residue W189 forms a face-to-edge type of $\pi - \pi$ stacking interaction with C7. Further, C7 forms hydrophobic contacts with residues G23, A138, L144, H163 and W189.

Phytochemical C8, picrasidine M, has a binding energy of -8.0 kcal/mol. C8 is a β -carboline alkaloid present in *Picrasma quassioides*. *Picrasma quassioides* has been reported to have antiviral and antifungal activities [50]. Figure 5.9H shows the cathepsin L residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with C8. The carboxylic group of residue D162 forms two C-H \cdots O type hydrogen bonds with C8. Also, residues M70 and G23 form hydrogen bonds of type C-H \cdots S and C-H \cdots O, respectively, with C8. Also, the residue W189 forms a face-to-edge type of $\pi - \pi$ stacking interaction with C8. Further, C8 forms hydrophobic contacts with residues G23, L69, D162 and W189.

Phytochemical C9, trisjuglone, has a binding energy of -8.0 kcal/mol. C9 is a naphthoquinone present in *Juglans regia* (i.e., walnut). In traditional medicine, *Juglans regia* has been reported to have anti-inflammatory, antifungal and antimicrobial activities [93]. Figure 5.9I shows the cathepsin L residues that form hydrogen bonds or $\pi - \pi$ stacking interactions with C9. The benzoquinone moiety present in C9 forms two C-H \cdots O type hydrogen bonds with residue Q21 and one N-H \cdots O type hydrogen bond with W189. In contrast, the other catalytic residue H163 forms a N-H \cdots O type hydrogen bond with C9. Also, the residue W189 forms a face-to-edge type of $\pi - \pi$ stacking interaction with C9. Further, C9 forms hydrophobic contacts with residues Q21, G23, C25, L144 and W189.

5.5 Virtual screening for SARS-CoV-2 Nsp13

In this section we describe the identification of potential phytochemical inhibitors of the SARS-CoV-2 helicase Nsp13 that can target the ATP binding site (Figure 5.10).

Similar to the workflow for the identification of potential inhibitors of host factors, we prepared the 10510 drug-like phytochemicals from Indian medicinal plants as described in Section 5.3.1. In the second stage, we prepared the SARS-CoV-2 Nsp13 for docking with prepared ligands. We have used the crystal structure of SARS-CoV-2 helicase Nsp13 (PDB 6ZSL) with 1.94 Å resolution from Protein Data Bank (PDB) for virtual screening. As of 19 February 2021, there were more than 50 crystal structures of SARS-CoV-2 Nsp13 in PDB. However, the crystal structure 6ZSL selected for this investigation has been validated by Wlodawer *et al.* [260], and furthermore, is the only SARS-CoV-2 Nsp13 structure not from Pan-Dataset Density Analysis (PanDDA). In the crystal structure 6ZSL for the SARS-CoV-2 Nsp13, we have gap-filled the structural coordinates for three missing amino acid residues (339–341) by aligning to the crystal structure for the SARS-CoV Nsp13 (PDB 6JYT). Subsequently, the gap-filled structure of SARS-CoV-2 Nsp13 was energy-minimized using UCSF Chimera [234].

Figure 5.3 shows the prepared crystal structure of SARS-CoV-2 Nsp13 wherein important residues in the ATP binding site have been highlighted [212, 213]. Previously, Jia *et al.* [212] had identified six residues, namely, K288, S289, D374, E375, Q404 and R567, in SARS-CoV Nsp13 to be crucial for ATP hydrolysis. These six key ATP binding site residues are also conserved in SARS-CoV-2 Nsp13. SARS-CoV-2 Nsp13 and SARS-CoV Nsp13 share 99.8% sequence identity with only one varying residue which is away from the ATP binding site and RNA-binding site [87, 213].

To capture the conformational diversity of the ATP binding site of the helicase Nsp13, a 100 ns MD simulation of the prepared crystal structure was performed to generate an ensemble of structures (Section 5.3.4). The MD simulation trajectory from 60 to 100 ns was subjected to geometric clustering using the Daura algorithm [261] based on the

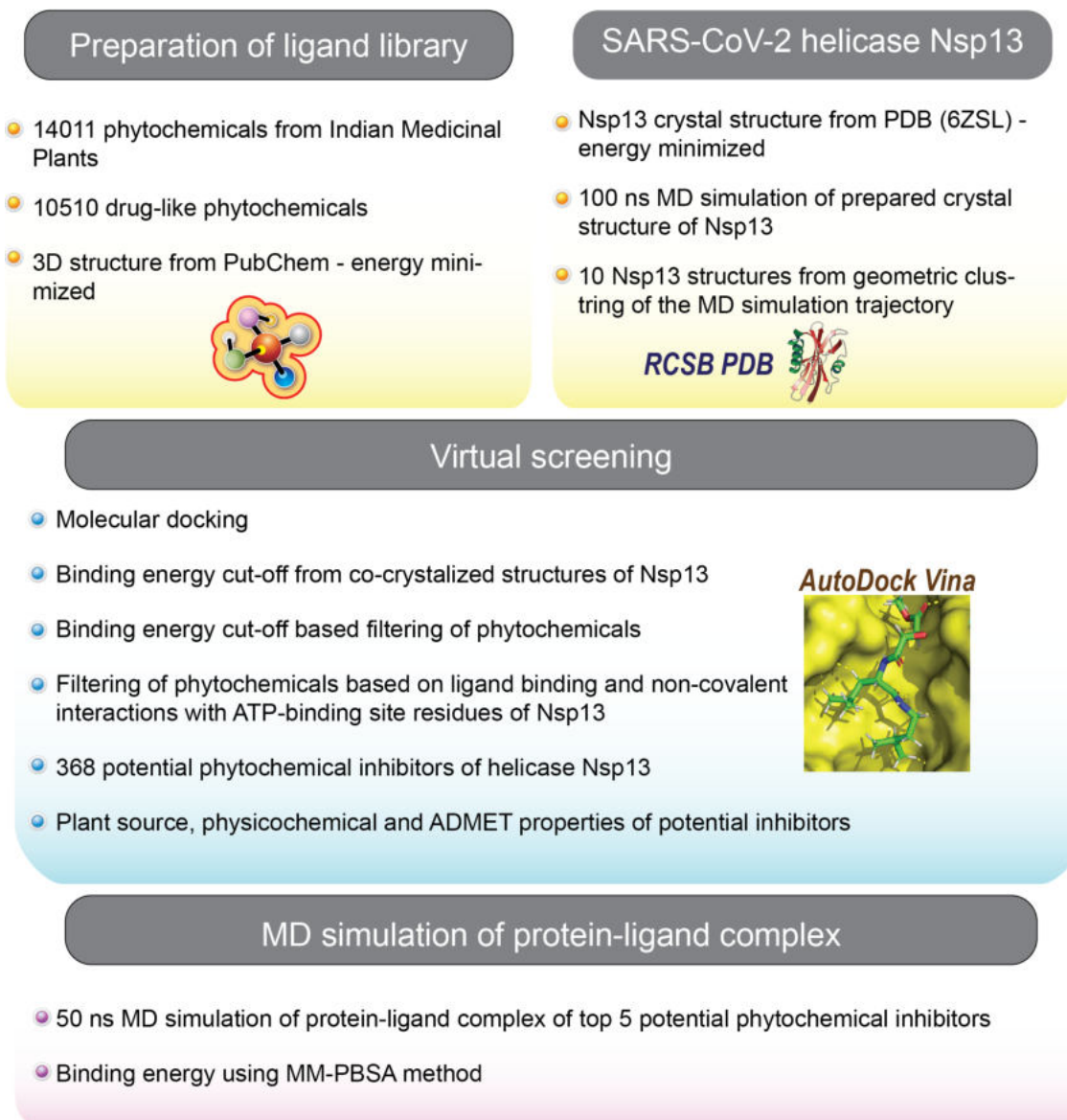


Figure 5.10: Workflow for the identification of potential phytochemical inhibitors of SARS-CoV-2 helicase Nsp13

conformations of the ATP binding site residues within 8 Å from the two phosphate ions in the prepared crystal structure of Nsp13. Using a clustering cut off of 1.2 Å, we were able to identify 23 clusters from the SARS-CoV-2 Nsp13 protein simulation, from which we selected ten representative mid-point structures corresponding to the top ten populated clusters accounting for 97.6% of the sampled conformations of the ATP binding site residues. Thus, along with the prepared crystal structure of SARS-CoV-2 Nsp13, ten structures based on clustering from the MD simulation trajectory of the protein were selected for virtual screening.

Next we performed protein-ligand docking using AutoDock Vina as described in Section 5.3.2. To decide on the binding energy cut off for identification of potential inhibitors of Nsp13, we used the known experimental information on binding from the twelve PanDDA co-crystallized structures of SARS-CoV-2 Nsp13 with ligands bound to the ATP binding site (PDB 5RM2, 5RM7, 5RLW, 5RL9, 5RLO, 5RLY, 5RLJ, 5RLI, 5RLV, 5RLR, 5RLN and 5RLS). The binding energy of the co-crystallized ligand in the twelve PanDDA co-crystallized structures of SARS-CoV-2 Nsp13 was estimated using score_only option of AutoDock Vina (Supplementary Table S5.8). We find the ligand UXG in the structure PDB 5RM2 has the best binding energy value of -5.14 kcal/mol. Thus, binding energy cut off of ≤ -5.14 kcal/mol was used as a first stage filter to identify potential inhibitors of Nsp13. Thereafter, molecular docking was performed for the prepared ligands in the phytochemical library against the Nsp13. For Nsp13, prepared crystal structure along with the ten structures (10+1) based on clustering from the MD simulation trajectory were used in molecular docking. We then filtered the subset of 5260 phytochemicals whose binding energy in the best docked pose with Nsp13 structures (10+1) is ≤ -5.14 kcal/mol.

For Nsp13, six key residues K288, S289, D374, E375, Q404 and R567 are crucial for ATP hydrolysis [212], therefore a potent inhibitor should either bind to or form non-covalent interactions with all the above six residues. Hence, we further used the ligand binding site residues and non-covalent interactions with the prepared crystal structure of Nsp13 as a second stage filter. Specifically, we checked that the phytochemical either

binds to or forms non-covalent interactions with all six key ATP binding site residues of the prepared crystal structure of Nsp13. At the end of the second stage of filtering, we obtained 368 phytochemicals (H1–H368) as potential inhibitors of SARS-CoV-2 Nsp13 targeting the crucial ATP binding site (Supplementary Table S5.9). Similar to the herbal source information provided for the phytochemical inhibitors of the host factors, for the 368 phytochemicals (H1 - H368) identified as potential inhibitors of Nsp13, we provide the Indian medicinal plant source and the information on potential antiviral use of the corresponding Indian medicinal plant in traditional medicine (Supplementary Table S5.10). In Supplementary Table S5.11 we provide the list of ligand binding site residues and non-covalent protein-ligand interactions for the identified phytochemical inhibitors of Nsp13. We also provide the physicochemical and predicted ADMET properties of the identified phytochemical inhibitors of Nsp13 (Supplementary Table S5.12).

Lastly, for the top five inhibitors of Nsp13 namely, H1 (Picrasidine M), H2 ((+)-Epiexcelsin), H3 (Isorhoeadine), H4 (Euphorbetin) and H5 (Picrasidine N), their respective protein-ligand complexes were analyzed using MD simulation and their interaction binding energy was computed using MM-PBSA method. The results from the MD simulation and MM-PBSA method based binding energy computations are described in Appendix B.

5.5.1 Potential Phytochemical Inhibitors of SARS-CoV-2 Nsp13

For the 368 potential phytochemical inhibitors (H1–H368) of SARS-CoV-2 Nsp13 identified in this work, we provide the PubChem identifier, chemical name, IUPAC name and chemical structure in SMILES format in Supplementary Table S5.9. In this section, a detailed description of the top ten potential phytochemical inhibitors (H1–H10) that have protein–ligand docking binding energies < -8.5 kcal/mol with the prepared crystal structure of SARS-CoV-2 Nsp13 (Tables 5.3) are discussed. Table 5.3 also provides the list of Indian herbs which can produce these top ten potential phytochemical inhibitors of SARS-CoV-2 Nsp13. The two-dimensional (2D) chemical structures and the hydrogen

bond interactions in the protein–ligand complex of these top ten inhibitors are shown in Figure 5.11 and 5.12, respectively.

Phytochemicals H1 (Picrasidine M) and H5 (Picrasidine N) are dimeric β -carboline-type alkaloid produced by the herb *Picrasma quassioides*. The herb *Picrasma quassioides* has been reported to have antiviral, antifungal and antiparasitic activities [93, 262, 263]. Additionally, the β -carboline alkaloids from *Picrasma quassioides* have been experimentally found to inhibit the RNA replication of the plant pathogen Tobacco mosaic virus (TMV) [264]. From Figure 5.12A it is seen residue S289 forms 2 hydrogen bonds with H1 (C–H \cdots O type and C–H \cdots N type), and residues Q404 and R567 form 1 N–H \cdots O type hydrogen bond each with H1. In case of H5, residues S289, Q404 and R567, which are among the six key ATP binding site residues, form 1 C–H \cdots O type, 1 N–H \cdots O type and 2 N–H \cdots O type hydrogen bonds with H5, respectively (Figure 5.12E).

Phytochemical H2 ((+)-Epiexcelsin) is a lignan produced by the herb *Litsea verticillata* [265]. Extract of *Litsea verticillata* has been experimentally found to have antiviral and anti-HIV activities [266]. From Figure 5.12B it is seen, the residues K288 and S289 form 2 hydrogen bonds (C–H \cdots O type and C–H \cdots N type) and 1 hydrogen bond (N–H \cdots O type) with H2, respectively.

Phytochemical H3 (Isorhoeadine) is a rhoeadine alkaloid produced by the herb *Papaver rhoeas* [93]. H3 forms 3 C–H \cdots O type hydrogen bonds with S289, E375 and Q404, 2 C–H \cdots N type hydrogen bonds with R443 and R567, and 2 N–H \cdots O type hydrogen bonds with K288 and Q404 (Figure 5.12C).

Phytochemical H4 (Euphorbetin) is a bicoumarin produced by the herb *Euphorbia lathyris* [267]. Figure 5.12D shows the extensive hydrogen bond network between H4 and the protein residues. Residues K288, S289, D374, E375 and R567 form 2 hydrogen bonds (C–H \cdots O type and N–H \cdots O type), 4 hydrogen bonds (2 of N–H \cdots O type, 1 of C–H \cdots O type and 1 of O–H \cdots O type), 1 hydrogen bond (O–H \cdots O type), 1 hydrogen bond (C–H \cdots O type) and 2 hydrogen bonds (N–H \cdots O type) with H4, respectively.

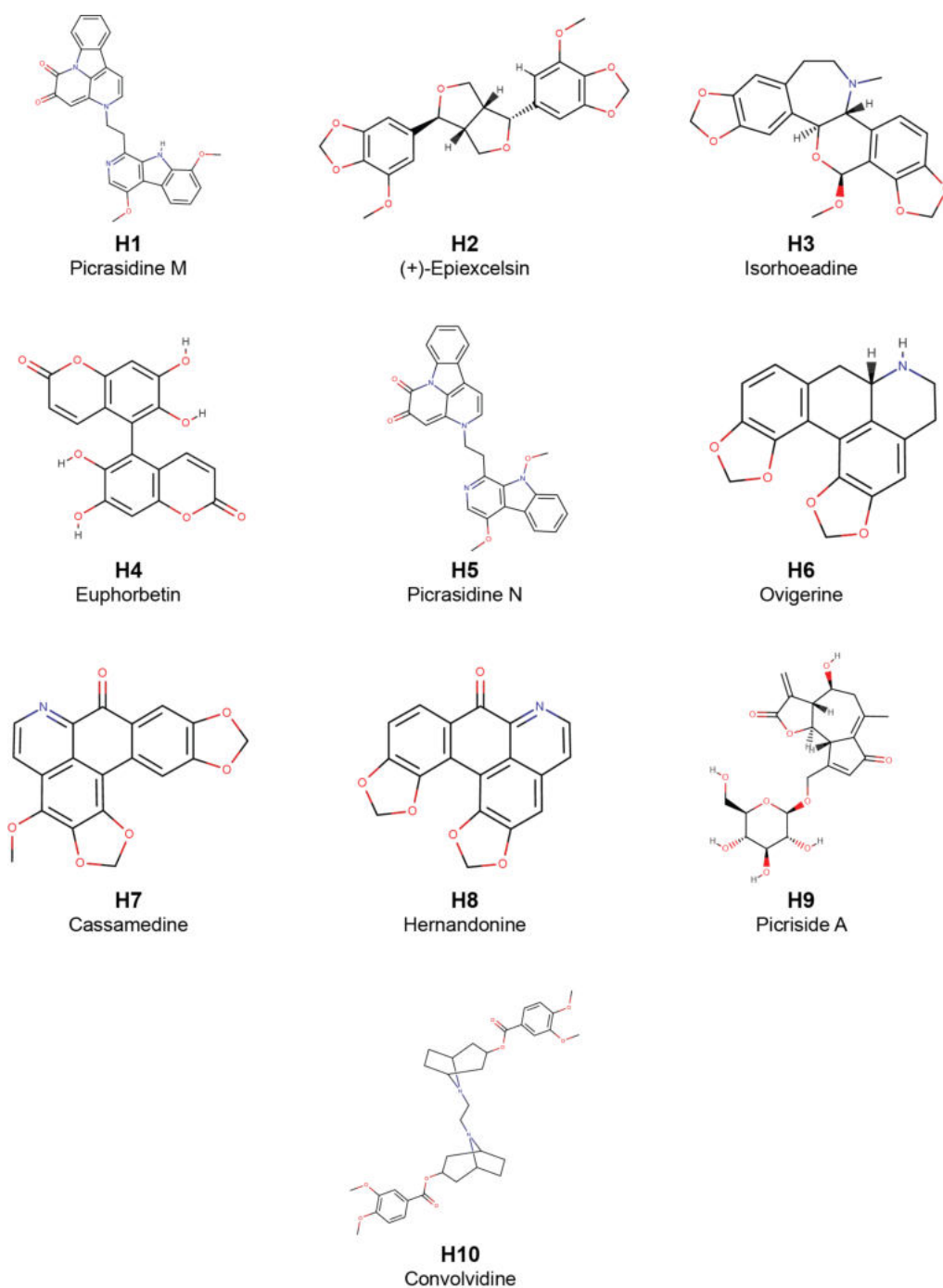
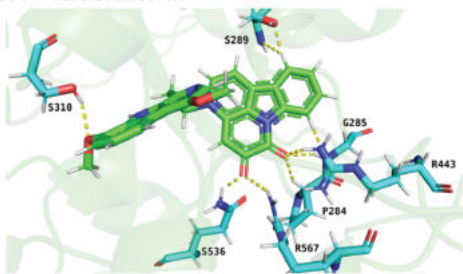
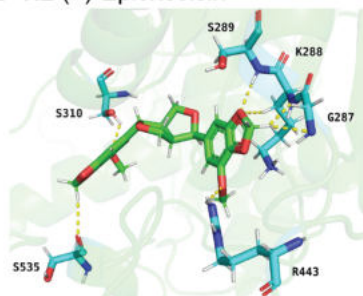


Figure 5.11: Molecular structure and chemical name of the top 10 phytochemical inhibitors (compounds H1–H10) of SARS-CoV-2 Nsp13.

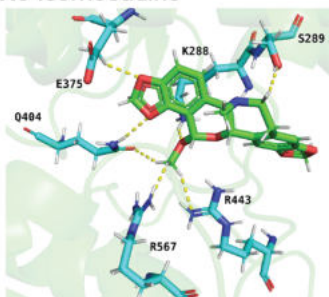
A H1 Picrasidine M



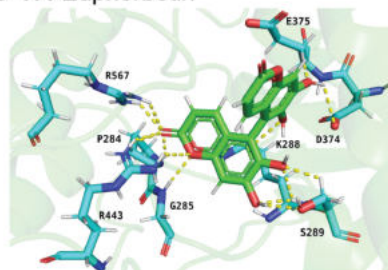
B H2 (+)-Epiexcelsin



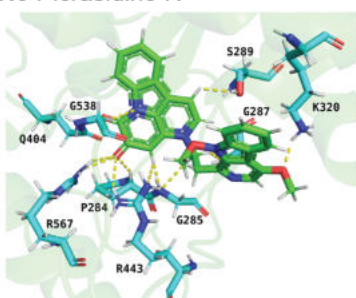
C H3 Isorhoeadine



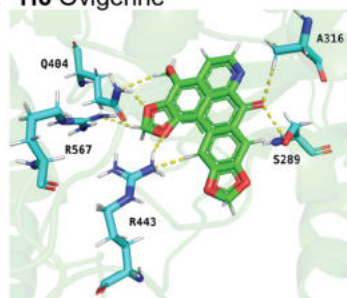
D H4 Euphorbetin



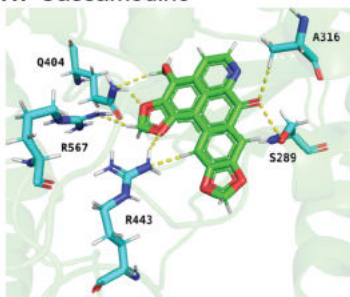
E H5 Picrasidine N



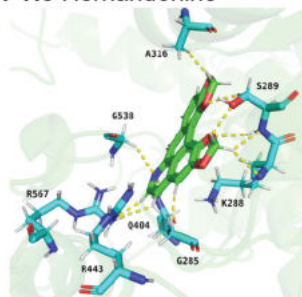
F H6 Ovigerine



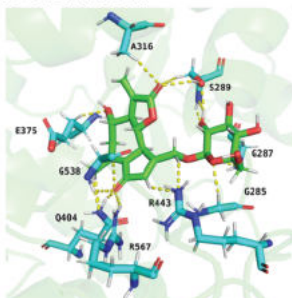
G H7 Cassamedine



H H8 Hernandonine



I H9 Picriside A



J H10 Convolvidine

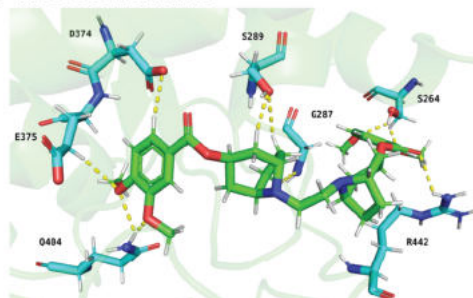


Figure 5.12 (previous page): Cartoon representation of the hydrogen bond interactions in the best-docked pose of the top 10 potential phytochemical inhibitors of SARS-CoV-2 helicase Nsp13. In this figure, hydrogen bond interactions are shown as yellow colored dotted lines between the residues of Nsp13 and the atoms of (A) H1, (B) H2, (C) H3, (D) H4, (E) H5, (F) H6, (G) H7, (H) H8, (I) H9 and (J) H10. The carbon atoms of the ligand are colored in green and the carbon atoms of the residues in Nsp13 are colored in cyan. The Nsp13 residues forming hydrogen bond interactions with the ligand are labeled with their one letter amino acid code and their residue number.

Phytochemicals H6 (Ovigerine) and H8 (Hernandonine) are both produced by the herbs *Hernandia guianensis* and *Hernandia nymphaeifolia*. Figure 5.12F shows the hydrogen bonds H6 forms with residues of Nsp13. In case of H8, residues K288, S289, Q404, and R567 form 1 hydrogen bond (C–H···N type), 4 hydrogen bonds (2 of C–H···O type, 1 of N–H···O type and 1 of C–H···N type), 1 hydrogen bond (N–H···N type) and 1 hydrogen bond (C–H···N type) with H8, respectively (Figure 5.12H).

Phytochemical H7 (Cassamedine) is an oxoaporphine alkaloid produced by the herb *Cassytha filiformis* [93, 262]. From Figure 5.12G it is seen, the residues S289, Q404 and R567 form 1 hydrogen bond (O–H···O type), 2 hydrogen bonds (N–H···O type and C–H···N type) and 1 hydrogen bond (C–H···N type) with H7, respectively.

Phytochemical H9 (Picriside A) is a glycoside produced by the herb *Picris hieracioides* [268]. H9 forms an extensive network of 17 hydrogen bonds with the residues of Nsp13 (Figure 5.12I).

Phytochemical H10 (Convolvidine) is a tropane alkaloid produced by the herb *Convolvulus prostratus* [269]. Residues S289, D374, E375 and Q404 form 2 hydrogen bonds (C–H···O type), 1 hydrogen bond (C–H···O type), 1 hydrogen bond (C–H···O type) and 2 hydrogen bonds (N–H···O type) with H10, respectively (Figure 5.12J).

5.6 Discussion

The current COVID-19 pandemic is a serious threat to humankind. More than 6 million people have died due to COVID-19 which is likely to become endemic past the pan-

demic phase. Computational approaches can be used to accelerate the identification and development of anti-COVID drugs. In this direction, protein-ligand docking and MD simulation are powerful computational methods to expedite the search for anti-COVID drugs by rapid identification of promising lead molecules. Here, we have used molecular docking and MD simulations in the search of natural compound inhibitors of: (a) two human proteases, TMPRSS2 and cathepsin L, that are key host factors in SARS-CoV-2 infection [85, 86, 207], and (b) SARS-Cov-2 helicase Nsp13 which is one of the key components of the viral replication and translation complex [87, 194].

Since early civilization, humans have used medicinal plants in different systems of traditional medicine to treat various ailments [270]. Specifically, traditional systems of Indian medicine including Ayurveda, Siddha and Unani have over centuries acquired invaluable knowledge on medicinal plants spanning the rich biodiversity of the subcontinent for treating various ailments including viral infections [50]. As plant-based natural products have been an indomitable source of lead molecules in the course of modern drug discovery [271], it is worthwhile to search for potential anti-COVID drugs among phytochemicals of Indian medicinal plants. Thus, in this work, we have performed virtual screening of 14011 phytochemicals that are produced by Indian medicinal plants to identify potential inhibitors of: (a) key host factors, TMPRSS2 and cathepsin L, and (b) SARS-Cov-2 Nsp13. For the identified top inhibitors, we have performed MD simulation, and thereafter, employed MM-PBSA method to compute binding energies of the protein-ligand complexes.

In our work presented in this chapter we have predicted 96 potential phytochemical inhibitors of TMPRSS2, 9 potential phytochemical inhibitors of cathepsin L and 368 potential phytochemicals inhibitors of SARS-CoV-2 Nsp13 targeting the ATP binding site. Furthermore, several herbal sources of the identified phytochemical inhibitors have been reported to have antiviral and/or anti-inflammatory use in traditional medicine (Table 5.1 and Table 5.2; Supplementary Tables S5.3 and S5.10). Additional *in vitro* and *in vivo* testing of the identified phytochemical inhibitors is needed before these molecules

can enter clinical trials against COVID-19. In conclusion, we expect the natural product inhibitors identified in this computational study will likely inform future research toward natural product-based anti-COVID therapeutics.

Supplementary Information

Supplementary Tables S5.1-S5.12 associated with this chapter are available for download from the GitHub repository: https://github.com/asamallab/PhDThesis-Vivek_Ananth_RP/blob/main/SI/ST_Chapter5.xlsx.

Phytochemical symbol	Docking binding energy (kcal/mol)	Chemical name	Plant source
T1	-9.6	Qingdainone	<i>Strobilanthes cusia</i> [*]
T2	-9.6	Edgeworoside C	Edgeworthia gardneri
T3	-9.6	Adlumidine	<i>Fumaria indica</i> [*]
T4	-9.3	Pseudo- α -Colubrine	<i>Strychnos nux-vomica</i> [*]
T5	-9.3	Bicuculline	<i>Fumaria indica</i> [*], <i>Corydalis govani</i> [*], <i>Nerium oleander</i> [*]
T6	-9.3	Strychnine N-oxide	<i>Strychnos nux-vomica</i> [*], <i>Strychnos ignatii</i> [*], <i>Strychnos colubrina</i> [*]
T7	-9.2	α -Colubrine	<i>Strychnos nux-vomica</i> [*], <i>Strychnos ignatii</i> [*], <i>Strychnos colubrina</i> [*]
T8	-9.2	Egenine	<i>Fumaria vaillantii</i> [*]
T9	-9.2	2-Hydroxy-3-methoxystrychnine	<i>Strychnos nux-vomica</i> [*]

Table 5.1: Herbal sources of the top 9 phytochemical inhibitors of TMPRSS2. The table provides the phytochemical symbol, docked binding energy (kcal/mol), chemical name and plant source for each phytochemical. Plant sources which have reported antiviral or anti-inflammatory use in traditional medicine literature are shown in bold and marked with an [*] sign.

Phytochemical symbol	Docking binding energy (kcal/mol)	Chemical name	Plant source
C1	-8.9	Arabinol	<i>Senna occidentalis</i> [*]
C2	-8.3	(+)-Oxoturkiyenine	Hypecoum pendulum
C3	-8.3	3Alpha,17Alpha-Cinchophylline	<i>Cinchona calisaya</i> [*]
C4	-8.2	Rugosanine B	<i>Ziziphus rugosa</i> [*]
C5	-8.2	Trichotomine	<i>Clerodendrum trichotomum</i> [*]
C6	-8.1	Tectol	<i>Tectona grandis</i> [*], <i>Tecomella undulata</i> [*]
C7	-8.1	Silymonin	<i>Silybum marianum</i> [*]
C8	-8	Picrasidine M	<i>Picrasma quassioides</i> [*]
C9	-8	Trisjuglone	<i>Juglans regia</i> [*]

Table 5.2: Herbal sources of top 9 phytochemical inhibitors of Cathepsin L. The table provides the phytochemical symbol, docked binding energy (kcal/mol), chemical name and plant source for each phytochemical. Plant sources which have reported antiviral or anti-inflammatory use in traditional medicine literature are shown in bold and marked with an [*] sign.

Phytochemical symbol	Docking binding energy (kcal/mol)	Chemical name	Plant source
H1	-10.2	Picrasidine M	<i>Picrasma quassioides</i> [*]
H2	-9	(+)-Epiexcelsin	Litsea verticillata
H3	-8.9	Isorhoeadine	<i>Papaver rhoeas</i> [*]
H4	-8.9	Euphorbetin	Euphorbia lathyris
H5	-8.9	Picrasidine N	<i>Picrasma quassioides</i> [*]
H6	-8.8	Ovigerine	Hernandia guianensis, Hernandia nymphaeifolia
H7	-8.8	Cassamedine	Cassytha filiformis
H8	-8.6	Hernandonine	Hernandia guianensis, Hernandia nymphaeifolia
H9	-8.6	Picriside A	Picris hieracioides
H10	-8.6	Convolvidine	<i>Convolvulus prostratus</i> [*]

Table 5.3: Herbal sources of top 10 phytochemical inhibitors of SARS-CoV-2 Nsp13. The table gives the phytochemical symbol, docked binding energy (kcal/mol), chemical name and plant source for each phytochemical. Plant sources which have reported antiviral use in traditional medicine literature are shown in bold and marked with an [*] sign.

Chapter 6

Summary and future outlook

In this thesis, we compiled, curated and explored the natural product spaces of Indian medicinal plants [50,51] and medicinal fungi [52] to enable traditional knowledge based drug discovery (Figure 6.1). Specifically, we describe the workflow for construction of IMPPAT (Indian Medicinal Plants, Phytochemistry And Therapeutics) versions 1.0 and 2.0, a curated database on Indian medicinal plants, their phytochemicals and their therapeutic uses. Subsequently, we exhaustively characterize the phytochemical space of the Indian medicinal plants captured in IMPPAT 2.0. Similar to IMPPAT, we constructed MeFSAT (Medicinal Fungi Secondary metabolite And Therapeutics), a curated digital database focused on secondary metabolites and therapeutic uses of medicinal fungi. We also characterized the secondary metabolite space of medicinal fungi captured in MeFSAT. The analysis of the natural product spaces from Indian medicinal plants and medicinal fungi reveals the unique features of these curated small molecule libraries. To further highlight the utility of the curated natural product space, we identified potential phytochemical inhibitors which can be used for development of anti-COVID drugs [53,54]. In the following section, we provide a brief summary of our research on multiple natural product spaces and their biological application toward new drug discovery presented across chapters of this thesis. We conclude with a discussion on future directions based on research presented in this thesis.

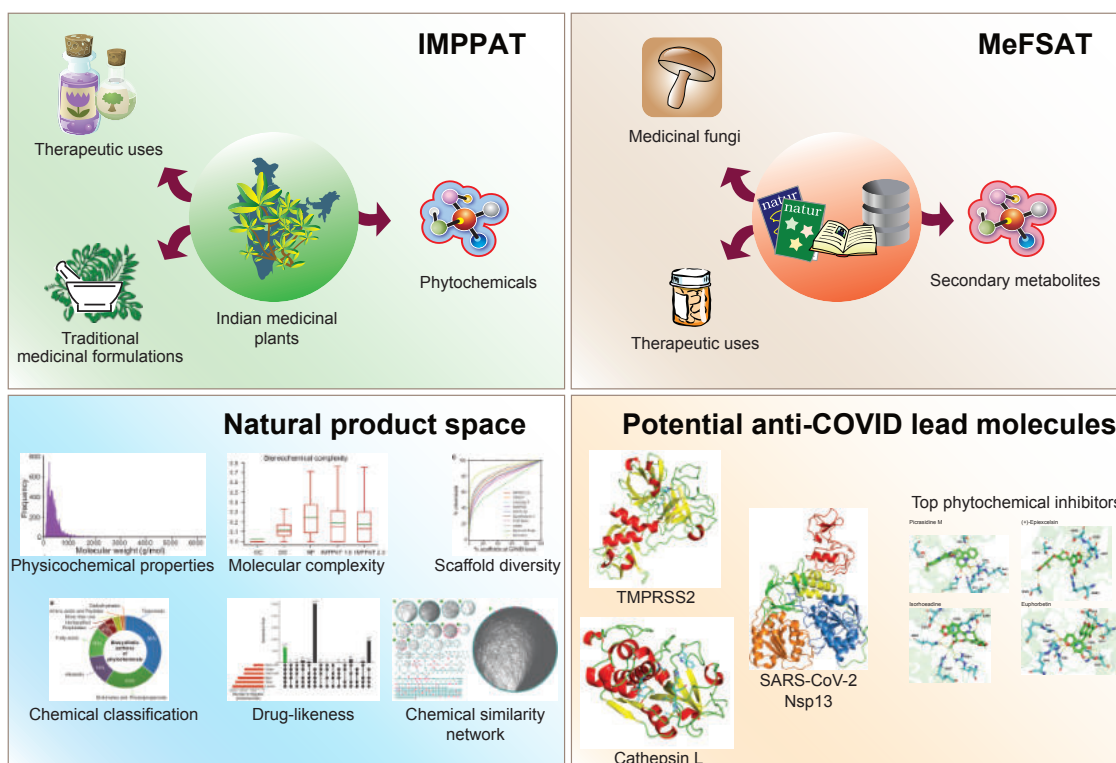


Figure 6.1: Summary of the research on compilation, curation and exploration of natural product spaces reported in this thesis.

6.1 Summary

Compilation, curation and exploration of the phytochemical space of Indian medicinal plants

IMPPAT is the largest manually curated resource to date on the phytochemicals from Indian medicinal plants. Chapter 2 presents the workflow for the construction of IMP-PAT version 1.0 (IMPPAT 1.0) and provides a detailed account on the update to create IMPPAT 2.0. IMPPAT 1.0 provided information on 1742 Indian Medicinal Plants, 9596 phytochemicals, and 1124 therapeutic uses. IMPPAT 1.0 also compiled chemical classification, chemical structures in different formats, physicochemical properties, drug-likeness scores, predicted absorption, distribution, metabolism, excretion and toxicity (ADMET) properties, and predicted human target proteins for phytochemicals in the database. IMP-PAT 1.0 also provided limited available information on the associations between Indian

medicinal plants and their use in traditional Indian medicinal formulations from Traditional knowledge digital library (TKDL; <http://www.tkd1.res.in>).

IMPPAT 2.0 is a significant enhancement and expansion over the previous version (IMPPAT 1.0; Table 2.2). The updated database is built upon the published data of earlier version IMPPAT 1.0, and provides information on 4010 Indian medicinal plants, 17967 phytochemicals, 1095 therapeutic uses and 1133 traditional Indian medicinal formulations. Figure 2.2 highlights the key features of IMPPAT 2.0. The coverage of the Indian medicinal plants in IMPPAT 2.0 is more than doubled in comparison with IMPPAT 1.0. Further, IMPPAT 2.0 now provides the phytochemical composition, therapeutic uses, and traditional medicinal formulations of Indian medicinal plants at the level of plant parts. Importantly, IMPPAT 2.0 provides a FAIR [60] compliant non-redundant *in silico* stereo-aware library of 17967 phytochemicals with their chemical structure information. The IMPPAT 2.0 phytochemical library is annotated with various features such as molecular scaffolds, predicted human target proteins, physicochemical properties, drug-likeness scores and predicted ADMET properties to facilitate its use in screening studies for identification of new drugs. The curated data on phytochemicals, therapeutic uses and traditional medicinal formulations of Indian medicinal plants in IMPPAT 2.0 can be accessed via user friendly web-interface at: <https://cb.imsc.res.in/imppat/>.

In chapter 3, we characterized the phytochemical space of Indian medicinal plants captured in IMPPAT 2.0. Specifically, we compared the molecular complexity and the molecular scaffold based structural diversity of the phytochemical space of IMPPAT 2.0 with other chemical libraries. We then assessed the drug-likeness of the phytochemicals using multiple drug-likeness scores. Lastly, we compared the overlap of the phytochemicals in IMPPAT 2.0 with phytochemicals from Chinese medicinal plants. IMPPAT 2.0 phytochemicals have high stereochemical and shape complexity similar to a representative natural product library from Clemons *et al.* [145]. Based on findings by Clemons *et al.* [84], who correlated the high values of stereochemical complexity and shape complexity of the natural product library with increased target protein specificity, the IMPPAT

2.0 phytochemicals are also more likely to be enriched with more specific protein binders than promiscuous binders. The structural diversity analysis using molecular scaffolds reveals that the IMPPAT 2.0 phytochemicals are structurally diverse with scaffold diversity similar to large natural product databases. We filtered 1335 drug-like phytochemicals which passed all the six drug-likeness scores considered in the analysis. We find only very few filtered drug-like phytochemicals are similar to any of the approved drugs. Also, the chemical similarity network of the drug-like phytochemicals highlighted the underlying structural diversity. The comparison of the IMPPAT 2.0 phytochemicals with the phytochemicals from Chinese medicinal plants reveals a minimal overlap between the two natural product libraries. These findings highlight the unique characteristics of the phytochemical space of Indian medicinal plants in IMPPAT 2.0.

Compilation, curation and exploration of the secondary metabolite space of medicinal fungi

Medicinal fungi have been used across the world in many traditional systems of medicine to treat human ailments for centuries. Before our work, there was no dedicated resource on secondary metabolites and therapeutic uses of medicinal fungi. To this end, we built MeFSAT a comprehensive manually curated database on secondary metabolites and therapeutic uses of medicinal fungi. Chapter 4 presents the workflow for compilation and curation of the secondary metabolite space of medicinal fungi captured in MeFSAT and the characterization of the associated chemical space. MeFSAT provides information on 184 medicinal fungi, 1830 secondary metabolites and 149 therapeutic uses from published literature.

The non-redundant *in silico* chemical library of 1830 secondary metabolites has been annotated with several features including 2D and 3D chemical structures, physicochemical properties, drug-likeness scores, predicted ADMET properties and predicted human target proteins. We find the secondary metabolites in MeFSAT have high stereochemical complexity and shape complexity close to the representative natural product library from

Clemons *et al.* [84]. This is similar to the finding for the IMPPAT phytochemicals, which also have high complexity values. This shows that the secondary metabolites of medicinal fungi in MeFSAT are also more likely to be enriched with specific protein binders than promiscuous binders. We filtered a subset of 228 drug-like secondary metabolites which pass all six drug-likeness scores considered in the analysis. The chemical similarity networks of all the secondary metabolites and the drug-like secondary metabolites in MeFSAT were both found to very sparse with several disconnected components and isolated nodes. This shows the underlying structural diversity among the secondary metabolites in MeFSAT. Also, we find only a small fraction of the secondary metabolites to be similar to any of the approved drugs. This highlights the potential utility of the secondary metabolites in MeFSAT for identification of new drugs. The compiled information on secondary metabolites and therapeutic uses of the medicinal fungi captured in MeFSAT can be accessed at: <https://cb.imsc.res.in/mefsat/>.

Identification of potential anti-COVID drugs

The COVID-19 pandemic caused by a novel betacoronavirus named SARS-CoV-2 has caused an unparalleled public health and economic emergency not witnessed in nearly a century. Identification and development of antiviral drugs against SARS-CoV-2 can aid in the ongoing efforts to combat COVID-19. Natural products from diverse organisms can be screened to identify potential anti-COVID drugs. Chapter 5 presents the biological application of the curated phytochemical space of Indian medicinal plants captured in IMPPAT for the identification of potential anti-COVID drugs. We focus on (a) two key host factors Transmembrane Protease Serine 2 (TMPRSS2) and cathepsin L, which play an important role in SARS-CoV-2 host cell entry [85, 86] and (b) SARS-CoV-2 helicase Nsp13 which unwinds the viral RNA helices in an ATP dependent manner playing an important role in viral life cycle [211] as targets to identify potential phytochemical inhibitors. We first use molecular docking based protein-ligand binding energy and then the ligand binding site residues and non-covalent interactions between protein and lig-

and to filter and identify phytochemical inhibitors that either bind to or form interactions with residues important for the specificity/activity of the target proteins. Altogether, we identified 96 inhibitors of TMPRSS2 and 9 inhibitors of cathepsin L among phytochemicals of Indian medicinal plants. Similarly, we identified 368 phytochemicals as potential inhibitors of SARS-CoV-2 helicase Nsp13. Lastly, we also provide the herbal sources of the above identified potential inhibitors of TMPRSS2, cathepsin L and SARS-CoV-2 Nsp13. The identified potential inhibitors can be taken up for further optimization to develop anti-COVID drugs after *in vitro* and *in vivo* experiments.

6.2 Future outlook

Natural products are one among the top sources for US FDA approved drugs [18]. Expanding the known natural product space through characterization of hitherto unexplored natural sources is an active area of research with potential for discovery of novel therapeutic molecules. In this thesis, we digitized more than 100 books on traditional Indian medicine to build IMPPAT, the largest to date phytochemical atlas of Indian medicinal plants. Latest version of IMPPAT provides a non-redundant *in silico* stereo-aware library of 17967 phytochemicals with a host of features. Similarly, we have captured the secondary metabolite space of medicinal fungi used in several traditional systems of medicine. Though the research reported in this thesis leads to by far the most comprehensive resource on Indian medicinal plants, their phytochemicals, their therapeutic uses and their use in traditional medicinal formulations, there exists an immense body of knowledge yet to be compiled and curated. On similar lines, though the exploratory analysis of the natural product spaces presented in this thesis revealed the underlying characteristics of these chemical spaces, there is possibility for further analysis to study the biological applications of the presented natural product spaces. Here we discuss our vision on how to further enrich and make better use of the resources presented in this thesis.

Our efforts to digitize more books on traditional medicine continues, given the vast

body of literature that lies in non-digital format. Here we would like to emphasize one aspect of our resource IMPPAT, the Indian medicinal plants - traditional medicinal formulation associations, which requires additional compilation and curation effort. This is due to the limited number of openly available formulations from TKDL (<http://www.tkdل.res.in>) and significant manual effort needed to collect data from books in vernacular languages. Thus, in future update of IMPPAT, we propose to include more data on traditional medicinal formulations using Indian medicinal plants. Any dataset becomes more valuable with finer level of granularity at which the data is captured and presented. In the latest version of IMPPAT (IMPPAT 2.0), we have captured the associations for Indian medicinal plants at the level of plant part such as leaves or roots or stem. This level of granularity can aid in better identification and extraction of phytochemicals from plants, and use of appropriate plant parts in the development of new botanical drugs. Another level of granularity we intend to add to our natural product resources is the quantitative content of the phytochemicals from different parts of the plant. We believe this additional dimension can facilitate better use of the natural product spaces presented in this thesis.

The natural product spaces of Indian medicinal plants and medicinal fungi have far more biological relevance and applications than revealed by the computational analysis presented in this thesis. To uncover and realize the full potential biological applications of the natural products, we propose to implement the following approaches. We intend to compile and curate the biological targets and activities of the natural products, specifically for the phytochemicals from Indian medicinal plants. IMPPAT already provides limited information on the predicted human target proteins of the phytochemicals using STITCH database [55]. Yet, there are more than one ways to map the phytochemicals to their biological targets [272]. The most reliable means is to compile data from published literature on the biological targets of the phytochemicals. Given the vastness of the natural product space, it is unlikely to find data for all the phytochemicals. Thus, we need to employ computational approaches such as large scale virtual screening of the

phytochemicals against known protein target space to map the biological targets. Another computational approach is to make use of machine learning based methods to predict the potential biological targets [272, 273]. We intend to use one or more of the above approaches to map the biological targets of the natural product spaces presented in this thesis. Apart from mapping the biological protein targets, providing the quantitative measure of the biological activity in terms of IC_{50} or EC_{50} of the phytochemical can enable development of computational models to predict biological activities of less studied phytochemicals [274]. Thus, we intend to incorporate available data on biological activity of the phytochemicals in future updates to build predictive computational models.

Also, the standardized data on phytochemicals and therapeutic uses of Indian medicinal plants captured in IMPPAT, provides a platform for applying network based system analysis to infer hidden connection which can throw light on therapeutic activity of the phytochemicals. Further, using cheminformatic analysis we can explore the natural product spaces for activity cliffs [275] and filter biologically interesting subspaces useful for drug discovery. We believe the valuable natural product spaces of Indian medicinal plants and medicinal fungi presented in this thesis will have far reaching applications going beyond our above limited discussion. In conclusion, the natural product spaces presented in this thesis will serve as a catalyst to propel further research on traditional knowledge based drug discovery.

Appendix A

Analysis of top inhibitors predicted for key host factors in SARS-CoV-2 infection

A.1 Reference inhibitors of TMPRSS2 and Cathepsin L

In order to identify potent phytochemical inhibitors of target proteins, we compared the binding energy of the best docked pose of ligands with binding energies of the best docked pose of known inhibitors of TMPRSS2 and cathepsin L obtained from AutoDock Vina.

Recent experiments have shown that both camostat mesylate and nafamostat mesylate, which are approved for human use in Japan, can block the TMPRSS2-dependent cell entry of SARS-CoV-2 [85,210]. By docking these two inhibitors to TMPRSS2 using AutoDock Vina, the predicted binding energies of camostat and nafamostat was found to be -7.4 kcal/mol and -8.5 kcal/mol, respectively. Figure A.1A,B show the best docked poses of nafamostat and camostat with TMPRSS2, and it is seen that both molecules form hydrogen bonds with the substrate binding residue D435. Importantly, in comparison to camostat mesylate, nafamostat mesylate in a recent experiment was shown to inhibit the TMPRSS2-dependent cell entry with 15-fold higher efficiency and an EC₅₀ value in

lower nanomolar range [210], and thus, the docked binding energies of these two known inhibitors are in line with experiments.

Recent experiments have shown that the small molecules E-64d and PC-0626568 (SID26681509) can block the cathepsin L-dependent cell entry of SARS-CoV-2 [85, 86, 207]. Note that cathepsin L is one of 11 cysteine cathepsin proteases encoded by the human genome, and the cathepsins share a high sequence similarity to papain, a non-specific plant protease [276]. E64-d is a broad spectrum inhibitor which can inhibit proteases cathepsins B, H, L and calpain, while PC-0626568 is a specific inhibitor of cathepsin L [86]. Moreover, a recent study [86] used the specific inhibitor PC-0626568 of cathepsin L to conclude that cathepsin L rather than cathepsin B is important for cell entry of SARS-CoV-2. Along with the above-mentioned two inhibitors of cathepsin L, we have also considered the co-crystallized inhibitor GH4 present in the crystal structure of cathepsin L (PDB 5MQY) as another reference inhibitor of cathepsin L. We remark that while recent experiments [85, 86, 207] have shown that E-64d and PC-0626568 can block the cathepsin L-dependent cell entry of SARS-CoV-2, similar experimental data specific to SARS-CoV-2 infection is lacking for cathepsin L inhibitor GH4.

By docking these three known inhibitors to cathepsin L using AutoDock Vina, the predicted binding energies of E-64d, PC-0626568 and GH4 were found to be -5.0 kcal/mol, -8.0 kcal/mol and -6.3 kcal/mol, respectively. Notably, the docked binding energies are in line with known specificity of E-64d and PC-0626568 to cathepsin L. Figure A.1C-E show the best docked poses of PC-0626568, E-64d and GH4 with cathepsin L. It is seen that both E-64d and PC-0626568 form hydrogen bonds with both catalytic residues C25 and H163. We have compared the docked pose of GH4 obtained from AutoDock Vina with the pose of GH4 in the co-crystallized structure of cathepsin L, and the RMSD between the heavy atoms in the two poses was found to be 0.786 Å. Figure A.2 shows the superimposed structures of the docked pose of GH4 from AutoDock Vina and the pose of GH4 in the co-crystallized structure of cathepsin L.

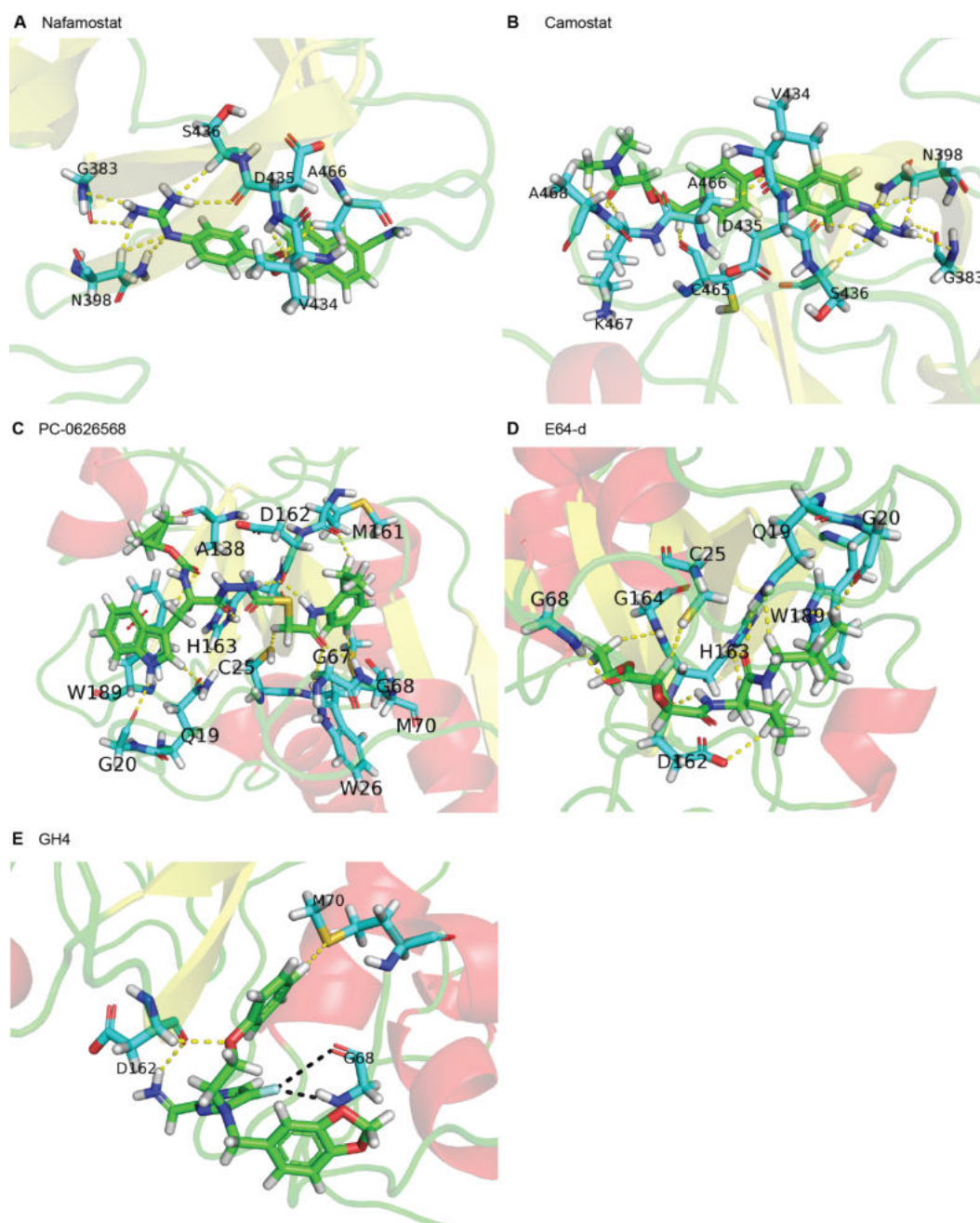


Figure A.1: Cartoon representation of the protein-ligand interactions of the known inhibitors of TMPRSS2 and cathepsin L. Interactions of TMPRSS2 residues with atoms of (A) Nafamostat, and (B) Camostat. Interactions of cathepsin L residues with atoms of (C) PC-0626568, (D) E-64d and (E) GH4. The carbon atoms of the ligand are shown in green colour while the carbon atoms of the amino acid residues in TMPRSS2 or cathepsin L are shown in cyan colour. TMPRSS2 or cathepsin L residues interacting with the ligand atoms via hydrogen bonds or $\pi - \pi$ stacking or halogen bonds are labelled with their corresponding one letter amino acid code along with their residue number in the protein sequence. The hydrogen bonds, $\pi - \pi$ stacking and halogen bonds are displayed using yellow, red and black dotted lines, respectively.

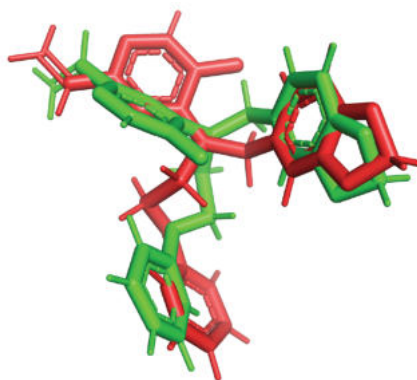


Figure A.2: Superimposition of the docked pose of GH4 with cathepsin L obtained from AutoDock Vina (shown in green colour) and the pose of GH4 in the co-crystallized structure with cathepsin L (shown in red colour).

A.2 Molecular Dynamics simulation of top inhibitors

In order to investigate the stability of the protein-ligand complexes of the identified inhibitors of TMPRSS2 and cathepsin L in Chapter 5, MD simulation of 180 ns was performed for top three inhibitors of TMPRSS2 namely, qingdainone (T1), edgeworoside C (T2) and adlumidine (T3), and of cathepsin L namely, ararobinol (C1), (+)-oxoturkiyenine (C2) and $3\alpha,17\alpha$ -cinchophylline (C3). Specifically, we have performed six 180 ns MD runs for protein-ligand complexes (TMPRSS2-T1, TMPRSS2-T2, TMPRSS2-T3, cathepsin L-C1, cathepsin L-C2 and cathepsin L-C3) and two 180 ns MD runs for uncomplexed TMPRSS2 and cathepsin L proteins. To assess the stability of the six protein-ligand complexes, we computed radius of gyration (R_g) of the protein, root mean square deviation (RMSD) of the C_α atoms of the protein, root mean square fluctuations (RMSF) of the C_α atoms of the protein, RMSD of the ligand and finally distance of the center of mass of the ligand from the center of mass of the catalytic residues or substrate binding residues of the protein in complex with the ligand (Figure A.3 and Figure A.4).

The R_g value of TMPRSS2 in complex with T1, T2 and T3 remains largely stable throughout the MD simulation (Figure A.3A). This implies that the top inhibitors of TMPRSS2 namely, T1, T2 and T3 do not induce any major structural changes to TMPRSS2 and TMPRSS2 remains structurally stable in complex with these inhibitors. TMPRSS2

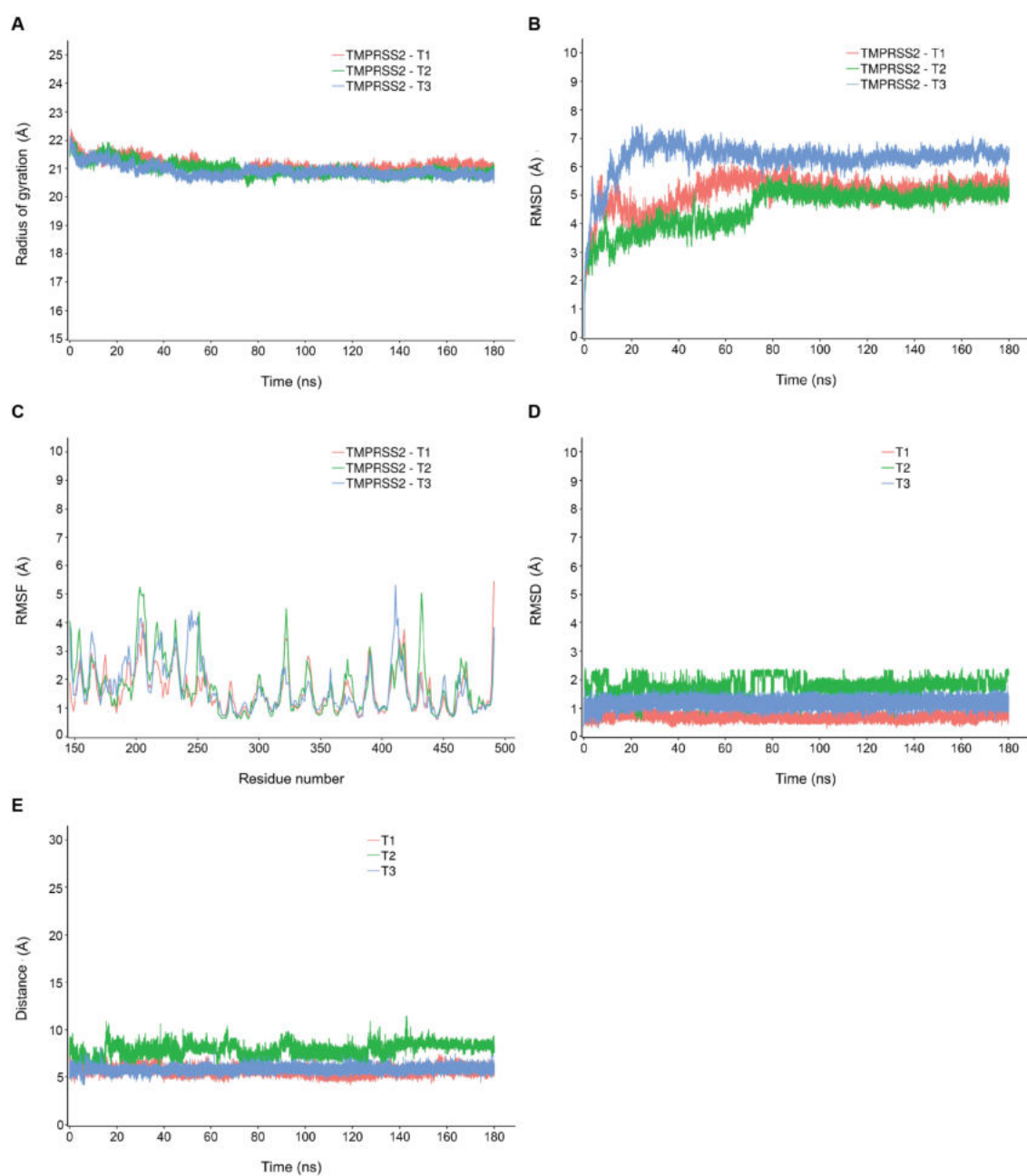


Figure A.3: Radius of gyration, RMSD, RMSF and distance from binding site residue derived from MD simulation of TMPRSS2-ligand complex. (A) Radius of gyration for TMPRSS2 in complex with T1, T2 and T3, (B) RMSD for TMPRSS2 in complex with T1, T2 and T3, (C) RMSF for TMPRSS2 in complex with T1, T2 and T3, (D) RMSD of T1, T2 and T3, and (E) Distance of the center of mass of T1, T2 and T3 from the substrate binding residue D435 in TMPRSS2.

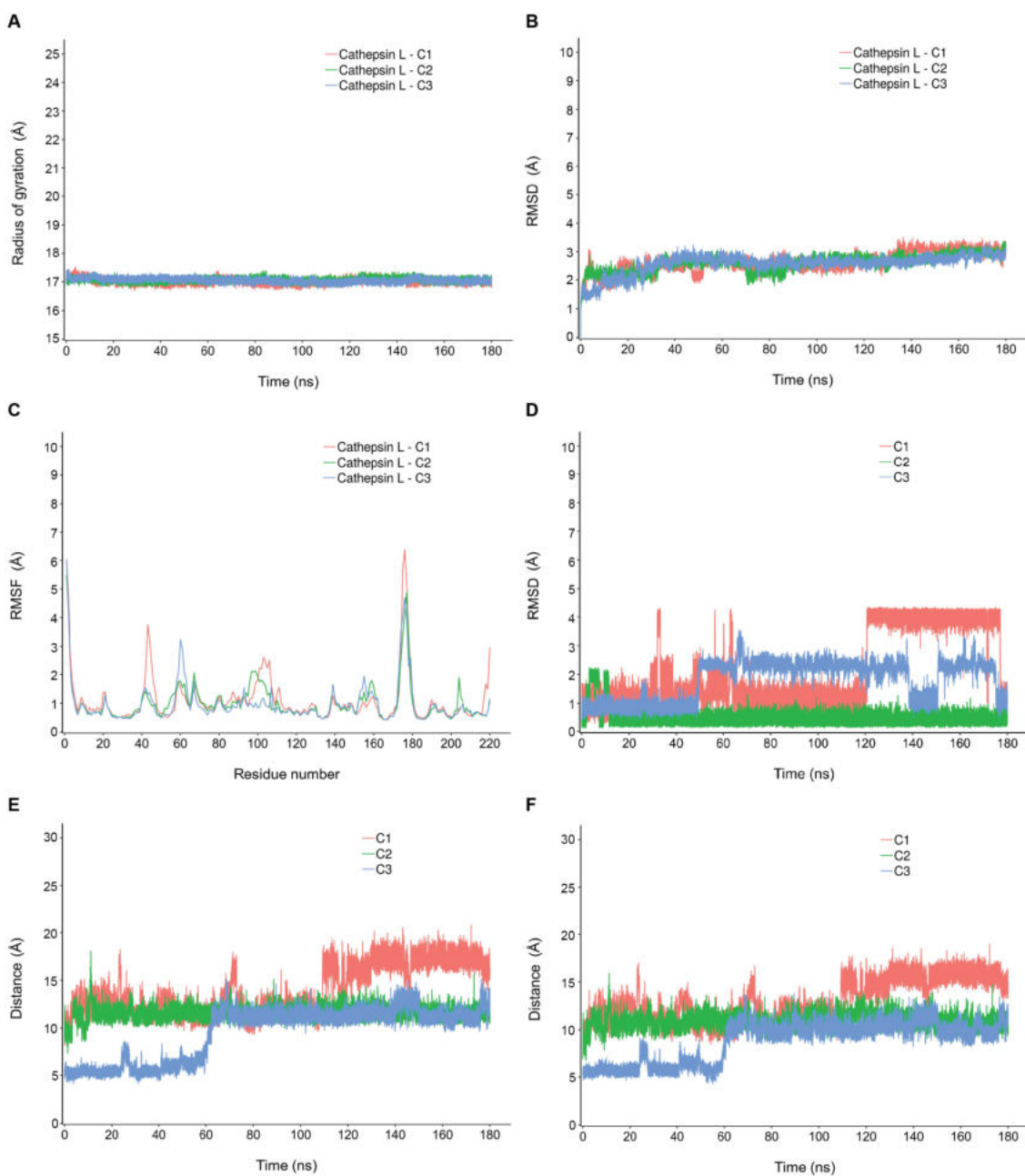


Figure A.4: Radius of gyration, RMSD, RMSF and distance from binding site residue derived from MD simulation of cathepsin L-ligand complex. (A) Radius of gyration for cathepsin L in complex with C1, C2 and C3, (B) RMSD for cathepsin L in complex with C1, C2 and C3, (C) RMSF for cathepsin L in complex with C1, C2 and C3, (D) RMSD of C1, C2 and C3, (E) Distance of the center of mass of C1, C2 and C3 from the catalytic residue C25 in cathepsin L, and (F) Distance of the center of mass of C1, C2 and C3 from the catalytic residue H163 in cathepsin L.

in complex with T1, T2 and T3 has an average R_g value of $21.10 \pm 0.22 \text{ \AA}$, $20.96 \pm 0.24 \text{ \AA}$ and $20.91 \pm 0.21 \text{ \AA}$, respectively. Further, RMSD value of the C_α atoms of TMPRSS2 in complex with T1, T2 and T3 become stable after 80 ns (Figure A.3B). Over the 80 ns to 180 ns time interval, TMPRSS2 in complex with T1, T2 and T3 has an average RMSD value of $5.25 \pm 0.23 \text{ \AA}$, $5.01 \pm 0.16 \text{ \AA}$ and $6.34 \pm 0.18 \text{ \AA}$, respectively. Lastly, Figure A.3C shows the RMSF value per residue for TMPRSS2 in complex with T1, T2 and T3. R_g , RMSD and RMSF values of TMPRSS2 in complex with T1, T2 and T3 closely follow R_g , RMSD and RMSF values of TMPRSS2 uncomplexed protein (Figure A.3A–C; Figure A.5A–C). Figure A.6A–C show the superimposed snapshots at 120 ns, 140 ns and 160 ns of TMPRSS2-T1, TMPRSS2-T2 and TMPRSS2-T3 complexes, respectively. To further quantify the stability of the inhibitors T1, T2 and T3 bound to TMPRSS2, we have computed the RMSD of T1, T2 and T3 (Figure A.3D) and distance of the center of mass of T1, T2 and T3 from the center of mass of the substrate binding residue D435 in TMPRSS2 (Figure A.3E). Both RMSD of T1, T2 and T3 bound with TMPRSS2 and distance of the center of mass of T1, T2 and T3 from the center of mass of the substrate binding residue D435 becomes largely stable after 100 ns of the MD simulation (Figure A.3D,E).

Also, R_g value of cathepsin L in complex with C1, C2 and C3 is stable throughout the MD simulation implying C1, C2 and C3 do not induce any major structural changes to cathepsin L and cathepsin L remains structurally stable in complex with these inhibitors (Figure A.4A). Cathepsin L in complex with C1, C2 and C3 has an average R_g value of $17.00 \pm 0.10 \text{ \AA}$, $17.06 \pm 0.07 \text{ \AA}$ and $17.05 \pm 0.08 \text{ \AA}$, respectively. Similarly, RMSD value of the C_α atoms of cathepsin L in complex with C1, C2 and C3 become largely stable after 80 ns (Figure A.4B). Over the 80 ns to 180 ns time interval, cathepsin L in complex with C1, C2 and C3 has an average RMSD value of $2.81 \pm 0.28 \text{ \AA}$, $2.76 \pm 0.22 \text{ \AA}$ and $2.70 \pm 0.14 \text{ \AA}$, respectively. Figure A.4C shows the RMSF value per residue for cathepsin L in complex with C1, C2 and C3. As in the case of TMPRSS2, R_g , RMSD and RMSF values of cathepsin L in complex with C1, C2 and C3 closely follow R_g , RMSD

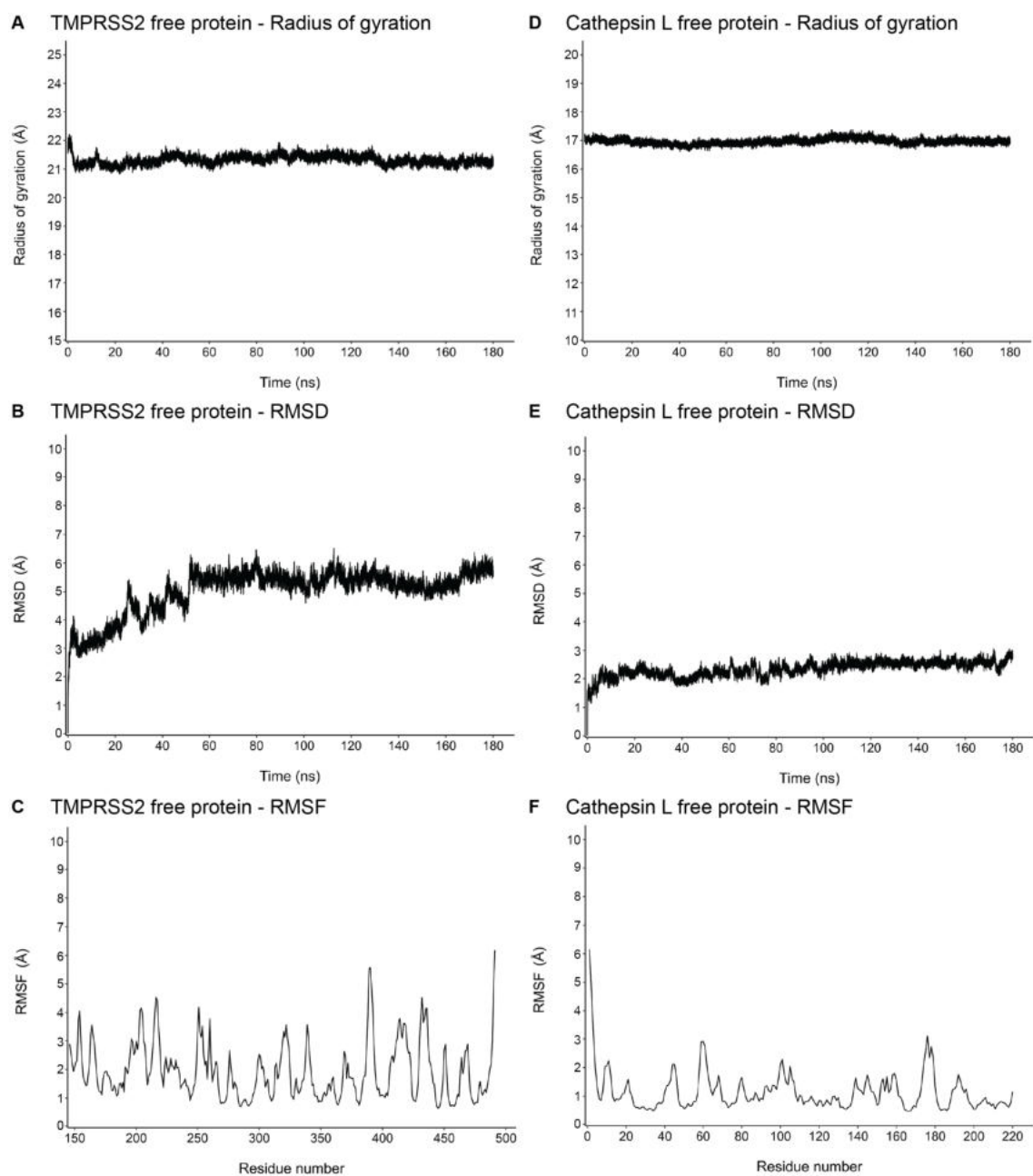
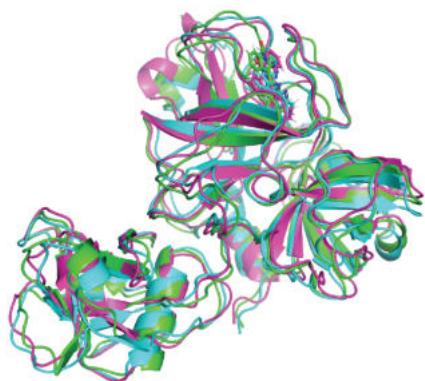
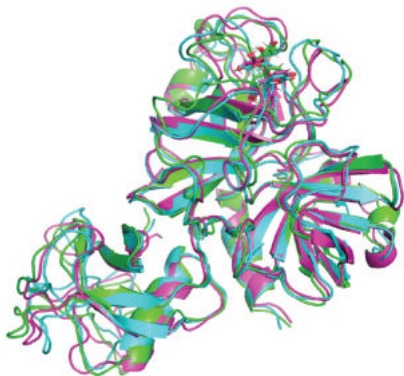


Figure A.5: Radius of gyration, RMSD and RMSF from MD simulation of uncomplexed TM-PRSS2 free protein and cathepsin L free protein. (A) Radius of gyration, (B) RMSD, and (C) RMSF for uncomplexed TMPRSS2 free protein. (D) Radius of gyration, (E) RMSD, and (F) RMSF for uncomplexed cathepsin L free protein.

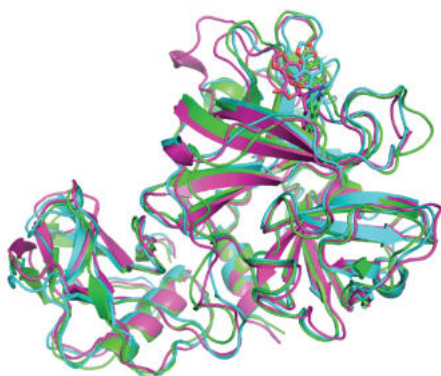
A TMPRSS2 - T1



B TMPRSS2 - T2



C TMPRSS2 - T3



D Cathepsin L - C1



E Cathepsin L - C2



F Cathepsin L - C3

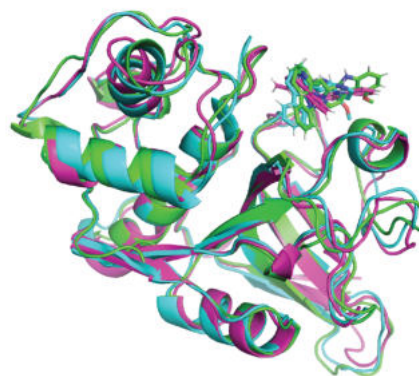


Figure A.6: Superimposition of the snapshots at 120 ns, 140 ns and 160 ns of (A) TMPRSS2-T1 complex, (B) TMPRSS2-T2 complex, (C) TMPRSS2-T3 complex, (D) cathepsin L-C1 complex, (E) cathepsin L-C2 complex and (F) cathepsin L-C3 complex obtained from their respective MD simulation trajectories.

and RMSF values of cathepsin L uncomplexed protein (Figure A.4A-C; Figure A.5D-F). Figure A.6D-F show the superimposed snapshots at 120 ns, 140 ns and 160 ns of cathepsin L-C1, cathepsin L-C2 and cathepsin L-C3 complexes, respectively. In order to quantify the stability of the inhibitors C1, C2 and C3 bound to cathepsin L we have also computed the RMSD of C1, C2 and C3 (Figure A.4D) and distance of the center of mass of C1, C2 and C3 from the center of mass of the catalytic residues C25 (Figure A.4E) and H163 (Figure A.4F) in cathepsin L. C1 has a largely stable RMSD after 120 ns of the MD simulation, C2 has the lowest and most stable RMSD in comparison with C1 and C3, and C3 shows a largely stable RMSD from 50 ns to 130 ns and from 150 ns to 170 ns of the MD simulation (Figure A.4D). Distance of the center of mass of C1, C2 and C3 from the center of mass of the catalytic residues C25 and H163 in cathepsin L also remains largely consistent after 120 ns of the MD simulation (Figure A.4E,F).

A.3 MM-PBSA binding energy of top inhibitors

Molecular Mechanics Poisson–Boltzmann Surface Area (MM-PBSA) is an extensively used method to compute the binding energy of small molecules with biological macromolecules such as proteins [277]. Notably, the protein-ligand binding energy computed using MM-PBSA method has been reported to be more accurate than that obtained from protein-ligand docking [277]. Thus, we have computed the binding energy for the top three inhibitors of TMPRSS2 and cathepsin L using MM-PBSA method. From the 180 ns MD simulation of the six protein-ligand complexes (TMPRSS2-T1, TMPRSS2-T2, TMPRSS2-T3, cathepsin L-C1, cathepsin L-C2 and cathepsin L-C3), 80 snapshots were obtained between 100 ns to 180 ns of the simulation at an interval of 1 ns along the trajectory, and thereafter, the 80 snapshots were used to compute the binding energy using `g_mmpbsa` [278,279]. The final binding energy is the sum of van der Waals, electrostatic, polar solvation, and solvent accessible surface area (SASA) energy components. The contribution of each of the above components to the binding energy of the top inhibitors is shown in Table A.1. While TMPRSS2-T1 complex has the lowest binding energy value

of -39.15 ± 2.799 kcal/mol, TMPRSS2-T2 and TMPRSS2-T3 complexes have binding energy value of -30.284 ± 3.585 kcal/mol and -27.386 ± 2.077 kcal/mol, respectively (Table A.1). In case of cathepsin L, cathepsin L-C1, cathepsin L-C2 and cathepsin L-C3 complexes have binding energy value of -22.384 ± 3.420 kcal/mol, -20.577 ± 3.600 kcal/mol and -26.156 ± 3.433 kcal/mol, respectively (Table A.1).

Protein-Ligand Complex	Binding Energy (kcal/mol)	Van Der Waals Energy (kcal/mol)	Electrostatic Energy (kcal/mol)	Polar Solvation Energy (kcal/mol)	SASA Energy (kcal/mol)
TMPRSS2-T1	-39.15 ± 2.799	-54.285 ± 2.903	-3.031 ± 1.439	22.844 ± 2.44	-4.678 ± 0.237
TMPRSS2-T2	-30.284 ± 3.585	-49.048 ± 3.838	-12.501 ± 4.884	35.978 ± 5.226	-4.712 ± 0.320
TMPRSS2-T3	-27.386 ± 2.077	-39.379 ± 2.109	-8.846 ± 1.423	24.359 ± 2.157	-3.52 ± 0.210
cathepsin L-C1	-22.384 ± 3.420	-25.296 ± 3.127	-2.214 ± 1.661	7.988 ± 4.103	-2.861 ± 0.366
cathepsin L-C2	-20.577 ± 3.600	-30.129 ± 3.154	-4.572 ± 2.138	16.891 ± 3.533	-2.767 ± 0.234
cathepsin L-C3	-26.156 ± 3.433	-37.165 ± 3.308	-2.093 ± 1.379	16.958 ± 4.513	-3.856 ± 0.319

Table A.1: MM-PBSA based binding energy for the top three inhibitors of TMPRSS2 and cathepsin L.

Appendix B

Analysis of top inhibitors predicted for SARS-CoV-2 Nsp13

B.1 Comparison with ligands co-crystallized with Nsp13

We have examined the ten PanDDA co-crystallized structures of SARS-CoV-2 Nsp13 with ligands bound near the ATP binding site, namely, PDB 5RM2, 5RM7, 5RLW, 5RL9, 5RLO, 5RLY, 5RLJ, 5RLV, 5RLN and 5RLS, in relation to the top five phytochemicals identified as potential inhibitors of SARS-CoV-2 Nsp13 in Chapter 5. The ligands from PanDDA co-crystallized structures PDB 5RLI and 5RLR were not considered for comparison as they were found to be structurally similar to the ligand present in PDB 5RLJ. The comparison of the binding modes of the top five phytochemical inhibitors of Nsp13 and the ligands from the PanDDA co-crystallized structures reveals distinct modes of binding for the top inhibitors that are different from the ligands of the PanDDA structures (Figure B.1). Specifically, the binding mode of the top phytochemical inhibitor H1 in the ATP binding site of Nsp13 doesn't overlap with that of the ligand NOE from the PanDDA structure PDB 5RM7, which has the highest 2D structural similarity with H1 (Figure B.1A,D). The above observations suggest the need for further experimental examination through co-crystallized structures and *in vitro* binding studies for the phytochemical in-

hibitors uncovered here.

B.2 Molecular Dynamics simulation of top inhibitors

We performed 50 ns MD simulations of the protein–ligand complexes of the top five phytochemical inhibitors of SARS-CoV-2 Nsp13 identified in Chapter 5, namely, Picrasidine M (H1), (+)-Epiexcelsin (H2), Isorhoeadine (H3), Euphorbetin (H4) and Picrasidine N (H5), and the structural characteristics of the complexes were evaluated (Figure B.2).

Figure B.2A shows the R_g of SARS-CoV-2 Nsp13 in complex with the top five inhibitors. R_g remains largely compact throughout the MD simulation of the five protein–ligand complexes (average R_g values are Nsp13-H1 = 27.84 ± 0.16 Å, Nsp13-H2 = 27.97 ± 0.15 Å, Nsp13-H3 = 27.88 ± 0.15 Å, Nsp13-H4 = 28.07 ± 0.17 Å, and Nsp13-H5 = 27.98 ± 0.17 Å). The RMSD value of the C_α atoms of SARS-CoV-2 Nsp13 in complex with the top five inhibitors stabilizes after 20 ns (Figure B.2B; average RMSD C_α over 20 ns–50 ns are Nsp13-H1 = 2.71 ± 0.22 Å, Nsp13-H2 = 3.82 ± 0.29 Å, Nsp13-H3 = 3.05 ± 0.23 Å, Nsp13-H4 = 4.01 ± 0.30 Å, and Nsp13-H5 = 2.97 ± 0.23 Å). In addition, the RMSF value per residue in the MD simulations of the SARS-CoV-2 Nsp13 in complex with the top five inhibitors closely follows the RMSF value per residue in the MD simulation of the SARS-CoV-2 Nsp13 uncomplexed protein (Figure B.2C; Figure B.3). The superimposed snapshots at 20 ns, 30 ns, 40 ns and 50 ns from the MD simulations of the protein–ligand complexes Nsp13-H1, Nsp13-H2, Nsp13-H3, Nsp13-H4 and Nsp13-H5 shows the conformational variations upon ligand binding (Figure B.4). The binding of the inhibitors to the protein is good as characterized by the RMSD of the top inhibitors (H1–H5) (Figure B.2D) and the distance between the center of masses of the inhibitors (H1–H5) and the six key ATP binding site residues namely, K288, S289, D374, E375, Q404 and R567 (Figure B.5).

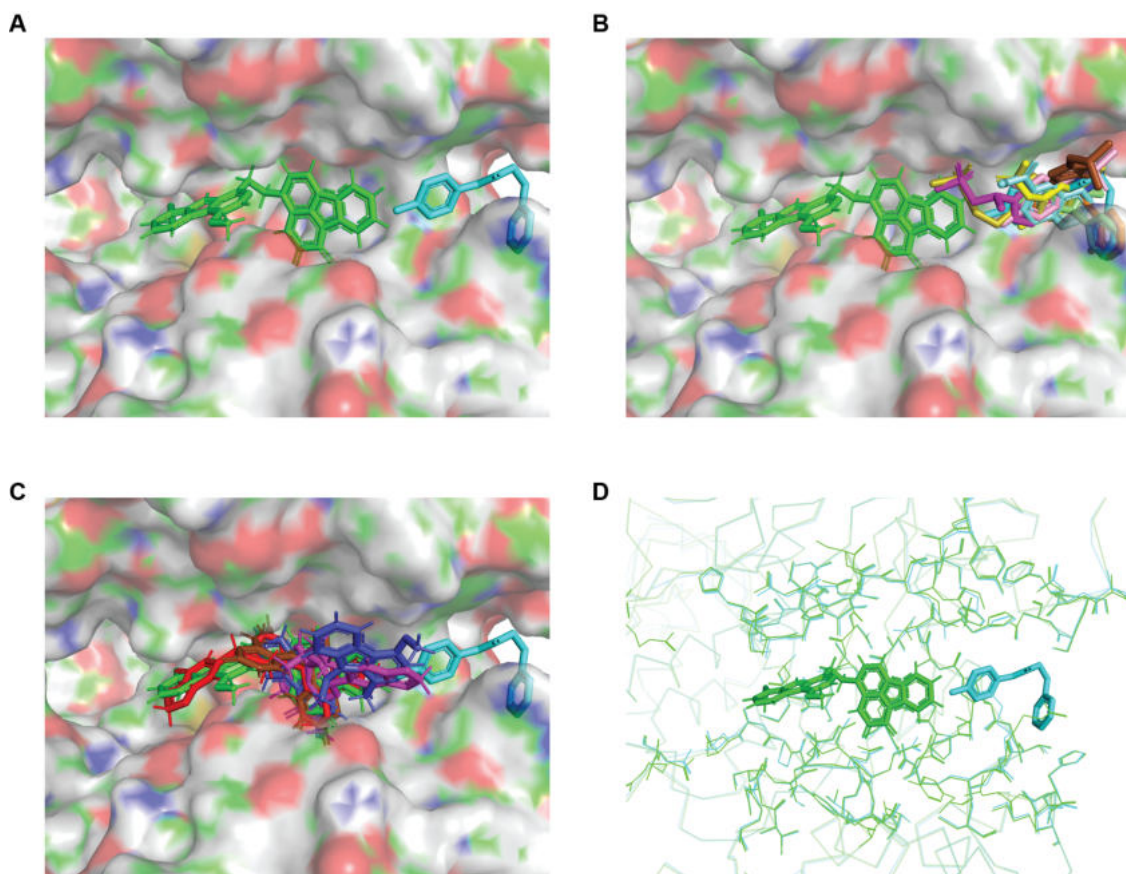


Figure B.1: Binding mode of the top inhibitors of SARS-CoV-2 Nsp13 and the ligands from PanDDA co-crystallized structures. (A) The top phytochemical inhibitor H1 (green) and the ligand NOE (cyan) from the PanDDA structure PDB 5RM7 which has the highest 2D structural similarity with the H1, in the pocket of Nsp13 crystal structure (PDB 6ZSL) visualized as surface. (B) The top phytochemical inhibitor H1 (green) and the ten ligands from the ten PanDDA structures in the pocket of Nsp13 crystal structure (PDB 6ZSL) visualized as surface. (C) The top five phytochemical inhibitors H1 (green), H2 (red), H3 (magenta), H4 (brown) and H5 (blue), and the ligand NOE (cyan) from the PanDDA structure PDB 5RM7 in the pocket of Nsp13 crystal structure (PDB 6ZSL) visualized as surface. (D) The top phytochemical inhibitor H1 (green), the ligand NOE (cyan) from the PanDDA crystal structure PDB 5RM7, with their respective protein structure backbone visualized as a ribbon. In (D), the protein atoms in the ligand binding site have been visualized as lines.

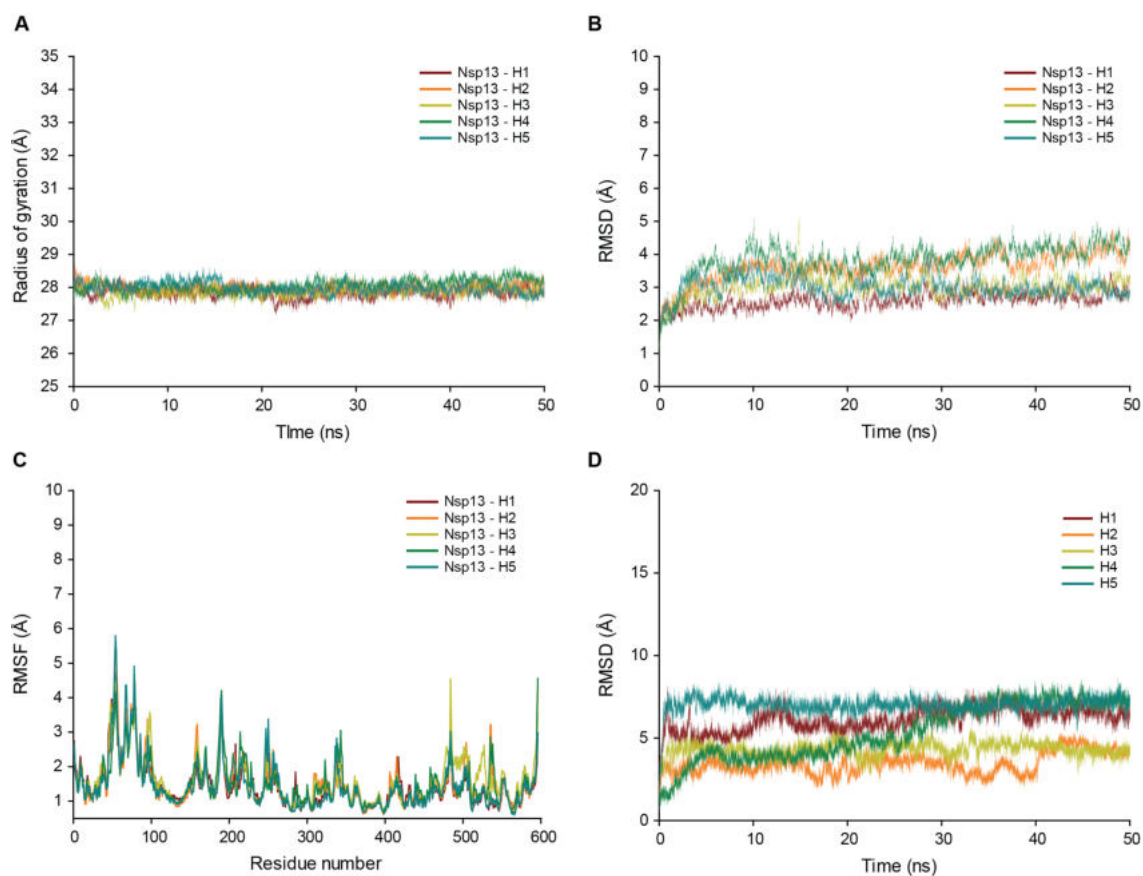


Figure B.2: Based on the 50 ns MD simulations of the protein–ligand complexes, the figure shows the (A) R_g , (B) RMSD and (C) RMSF of the SARS-CoV-2 Nsp13 in complex with the top five phytochemical inhibitors, namely, Nsp13-H1, Nsp13-H2, Nsp13-H3, Nsp13-H4 and Nsp13-H5, and (D) RMSD of the top five phytochemical inhibitors H1, H2, H3, H4 and H5.

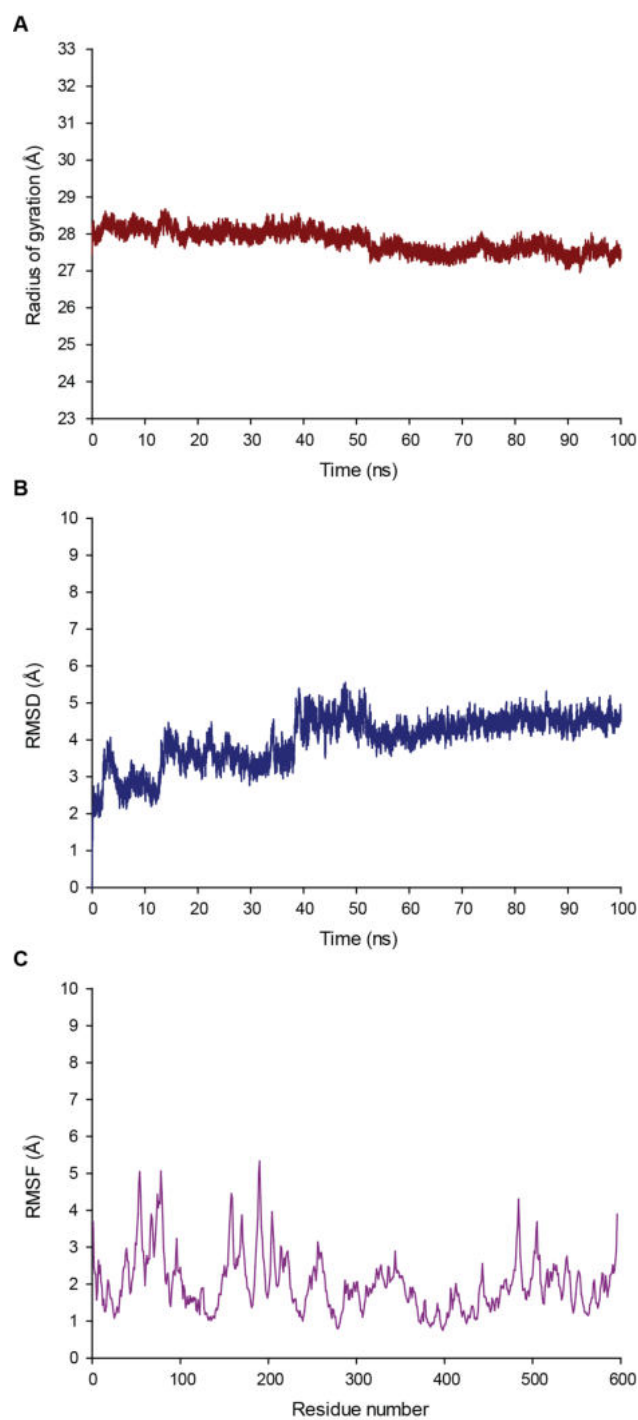


Figure B.3: Based on the 100 ns MD simulation of the uncomplexed protein, the figure shows the (A) Radius of gyration, (B) RMSD, and (C) RMSF of the prepared crystal structure of SARS-CoV-2 Nsp13.

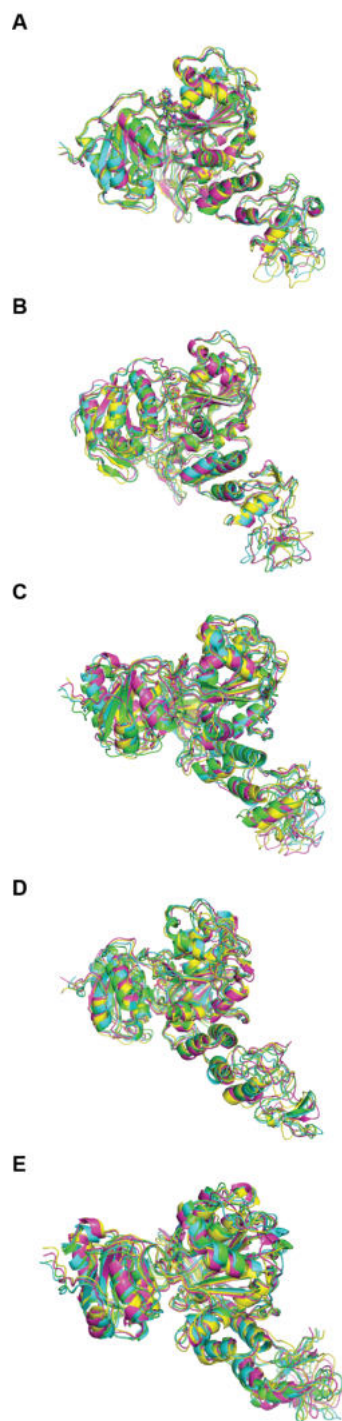


Figure B.4: Based on the 50 ns MD simulations of the protein-ligand complexes of top five inhibitors, the figure shows the superimposed snapshots at 20 ns, 30 ns, 40 ns and 50 ns of (A) Nsp13-H1 complex, (B) Nsp13-H2 complex, (C) Nsp13-H3 complex, (D) Nsp13-H4 complex and (E) Nsp13-H5 complex.

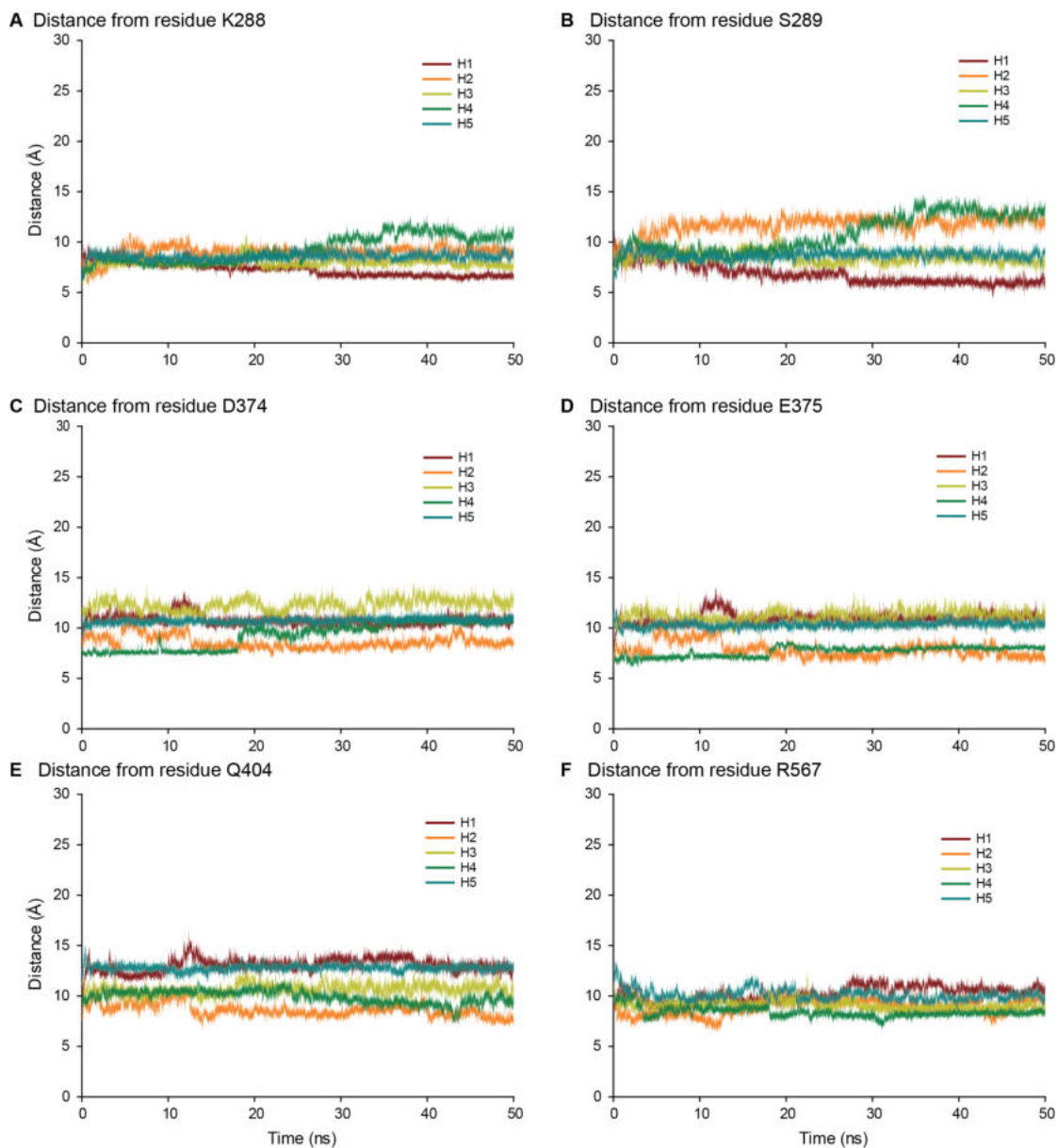


Figure B.5: Based on the MD simulations of the protein-ligand complexes of top five inhibitors, the figure shows the distance of the center of mass of the top five phytochemical inhibitors H1, H2, H3, H4 and H5 from (A) center of mass of residue K288, (B) center of mass of residue S289, (C) center of mass of residue D374, (D) center of mass of residue E375, (E) center of mass of residue Q404, and (F) center of mass of residue R567.

B.3 MM-PBSA binding energy of top inhibitors

The MM-PBSA based binding energy of the top five phytochemical inhibitors of SARS-CoV-2 Nsp13 (Table B.1) indicates the importance of the relative contributions of the van der Waals energy, the electrostatic energy, the polar solvation energy, and the solvent accessible surface area energy to the protein-ligand binding. The top five phytochemical inhibitors identified here, namely, Picrasidine M (H1), (+)-Epiexcelsin (H2), Isorhoeadine (H3), Euphorbetin (H4) and Picrasidine N (H5), have MM-PBSA based binding energy values of -13.211 ± 5.507 kcal/mol, -21.329 ± 4.067 kcal/mol, -17.618 ± 3.846 kcal/mol, -6.564 ± 5.422 kcal/mol and -11.76 ± 3.253 kcal/mol, respectively.

Protein-ligand complex	Binding energy (kcal/mol)	Van der waals energy (kcal/mol)	Electrostatic energy (kcal/mol)	Polar solvation energy (kcal/mol)	SASA energy (kcal/mol)
NSP13-H1	-13.211 ± 5.507	-52.006 ± 3.019	-14.428 ± 2.359	58.398 ± 6.429	-5.174 ± 0.270
NSP13-H2	-21.329 ± 4.067	-44.04 ± 2.424	-6.027 ± 1.608	33.337 ± 5.045	-4.599 ± 0.234
NSP13-H3	-17.618 ± 3.846	-44.531 ± 3.207	-5.494 ± 2.005	36.548 ± 4.424	-4.141 ± 0.320
NSP13-H4	-6.564 ± 5.422	-23.793 ± 4.342	-49.666 ± 7.567	70.915 ± 10.753	-4.02 ± 0.248
NSP13-H5	-11.76 ± 3.253	-47.293 ± 2.625	-13.453 ± 2.402	53.897 ± 4.958	-4.912 ± 0.334

Table B.1: MM-PBSA based binding energies for top five phytochemical inhibitors of SARS-CoV-2 Nsp13

References

- [1] Mora, C., Tittensor, D. P., Adl, S., Simpson, A. G. B. & Worm, B. How many species are there on earth and in the ocean? *PLOS Biology* **9**, 1–8 (2011).
- [2] Hoffmann, A. A. & Hercus, M. J. Environmental Stress as an Evolutionary Force. *BioScience* **50**, 217–226 (2000).
- [3] Voje, K. L., Holen, Ø. H., Liow, L. H. & Stenseth, N. C. The role of biotic forces in driving macroevolution: beyond the red queen. *Proceedings of the Royal Society B: Biological Sciences* **282**, 20150186 (2015).
- [4] Fiehn, O. Metabolomics—the link between genotypes and phenotypes. *Plant Molecular Biology* **48**, 155–171 (2002).
- [5] Wishart, D. S. Current Progress in computational metabolomics. *Briefings in Bioinformatics* **8**, 279–293 (2007).
- [6] Mosunova, O., Navarro-Muñoz, J. C. & Collemare, J. The biosynthesis of fungal secondary metabolites: From fundamentals to biotechnological applications. In Óscar Zaragoza & Casadevall, A. (eds.) *Encyclopedia of Mycology*, 458–476 (Elsevier, Oxford, 2021).
- [7] Zaynab, M. *et al.* Role of secondary metabolites in plant defense against pathogens. *Microbial Pathogenesis* **124**, 198–202 (2018).

- [8] Divekar, P. A. *et al.* Plant secondary metabolites as defense tools against herbivores for sustainable crop protection. *International Journal of Molecular Sciences* **23** (2022).
- [9] Sorokina, M. & Steinbeck, C. Review on natural products databases: where to find data in 2020. *Journal of Cheminformatics* **12**, 20 (2020).
- [10] Katiyar, C., Kanjilal, S., Gupta, A. & Katiyar, S. Drug discovery from plant sources: An integrated approach. *AYU (An International Quarterly Journal of Research in Ayurveda)* **33**, 10 (2012).
- [11] Gorse, A.-D. Diversity in medicinal chemistry space. *Current Topics in Medicinal Chemistry* **6**, 3–18 (2006).
- [12] Grigalunas, M., Brakmann, S. & Waldmann, H. Chemical evolution of natural product structure. *Journal of the American Chemical Society* **144**, 3314–3329 (2022).
- [13] Lipinski, C. & Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **432**, 855–861 (2004).
- [14] Saldívar-González, F. I. & Medina-Franco, J. L. Approaches for enhancing the analysis of chemical space for drug discovery. *Expert Opinion on Drug Discovery* 1–10 (2022).
- [15] Clayden, J., Greeves, N. & Warren, S. G. *Organic chemistry* (Oxford University Press, Oxford ; New York, 2012), 2nd edn.
- [16] Koehn, F. E. & Carter, G. T. The evolving role of natural products in drug discovery. *Nature Reviews Drug Discovery* **4**, 206–220 (2005).
- [17] Newman, D. J. & Cragg, G. M. Natural Products as Sources of New Drugs from 1981 to 2014. *Journal of Natural Products* **79**, 629–661 (2016).

- [18] Newman, D. J. & Cragg, G. M. Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of Natural Products* **83**, 770–803 (2020).
- [19] Li, J. W.-H. & Vederas, J. C. Drug Discovery and Natural Products: End of an Era or an Endless Frontier? *Science* **325**, 161–165 (2009).
- [20] Pye, C. R., Bertin, M. J., Lokey, R. S., Gerwick, W. H. & Linington, R. G. Retrospective analysis of natural products provides insights for future discovery trends. *Proceedings of the National Academy of Sciences* **114**, 5601–5606 (2017).
- [21] Lagunin, A. A. *et al.* Chemo- and bioinformatics resources for in silico drug discovery from medicinal plants beyond their traditional use: a critical review. *Natural Product Reports* **31**, 1585–1611 (2014).
- [22] Chen, Y. & Kirchmair, J. Cheminformatics in Natural Product-based Drug Discovery. *Molecular Informatics* **39**, 2000171 (2020).
- [23] Saldívar-González, F. I., Aldas-Bulos, V. D., Medina-Franco, J. L. & Plisson, F. Natural product drug discovery in the artificial intelligence era. *Chemical Science* **13**, 1526–1546 (2022).
- [24] Howes, M.-J. R. *et al.* Molecules from nature: Reconciling biodiversity conservation and global healthcare imperatives for sustainable use of medicinal plants and fungi. *PLANTS, PEOPLE, PLANET* **2**, 463–481 (2020).
- [25] Agrawal, D. C., Tsay, H.-S., Shyur, L.-F., Wu, Y.-C. & Wang, S.-Y. (eds.) *Medicinal Plants and Fungi: Recent Advances in Research and Development*, vol. 4 of *Medicinal and Aromatic Plants of the World* (Springer, Singapore, 2017).
- [26] Süntar, I. Importance of ethnopharmacological studies in drug discovery: role of medicinal plants. *Phytochemistry Reviews* **19**, 1199–1209 (2020).

- [27] Pandey, M. M., Rastogi, S. & Rawat, A. K. S. Indian traditional ayurvedic system of medicine and nutritional supplementation. *Evidence-Based Complementary and Alternative Medicine* **2013**, 376327 (2013).
- [28] Ravishankar, B. & Shukla, V. Indian Systems Of Medicine: A Brief Profile. *African Journal of Traditional, Complementary and Alternative Medicines* **4**, 319 (2008).
- [29] Jaiswal, Y. S. & Williams, L. L. A glimpse of ayurveda – the forgotten history and principles of indian traditional medicine. *Journal of Traditional and Complementary Medicine* **7**, 50–53 (2017).
- [30] Nadkarni, K. M. & Nadkarni, A. K. *Indian materia medica*. (Popular Book Depot, Bombay, 1955).
- [31] Dash, B. & Kashyap, L. *Materia Medica of Ayurveda* (Concept Publishing Company, 1999).
- [32] Gu, J., Gui, Y., Chen, L., Yuan, G. & Xu, X. CVDHD: a cardiovascular disease herbal database for drug discovery and network pharmacology. *Journal of Cheminformatics* **5**, 51 (2013).
- [33] Afendi, F. M. *et al.* KNAPSAcK Family Databases: Integrated Metabolite–Plant Species Databases for Multifaceted Plant Research. *Plant and Cell Physiology* **53**, e1 (2012).
- [34] Jensen, K., Panagiotou, G. & Kouskoumvekaki, I. Integrated Text Mining and Chemoinformatics Analysis Associates Diet to Health Benefit at Molecular Level. *PLOS Computational Biology* **10**, e1003432 (2014).
- [35] Jensen, K., Panagiotou, G. & Kouskoumvekaki, I. NutriChem: a systems chemical biology resource to explore the medicinal value of plant-based foods. *Nucleic Acids Research* **43**, D940–D945 (2015).

- [36] Pathania, S., Ramakrishnan, S. M. & Bagler, G. Phytochemica: a platform to explore phytochemicals of medicinal plants. *Database* **2015**, bav075 (2015).
- [37] Pathania, S., Ramakrishnan, S. M., Randhawa, V. & Bagler, G. SerpentinaDB: a database of plant-derived molecules of Rauvolfia serpentina. *BMC Complementary and Alternative Medicine* **15**, 262 (2015).
- [38] Zeng, X. *et al.* NPASS: natural product activity and species source database for natural product research, discovery and tool development. *Nucleic Acids Research* **46**, D1217–D1222 (2017).
- [39] Zeng, X. *et al.* CMAUP: a database of collective molecular activities of useful plants. *Nucleic Acids Research* **47**, D1118–D1127 (2018).
- [40] Yan, D. *et al.* HIT 2.0: an enhanced platform for Herbal Ingredients' Targets. *Nucleic Acids Research* **50**, D1238–D1243 (2021).
- [41] Chen, C. Y.-C. TCM Database@Taiwan: The World's Largest Traditional Chinese Medicine Database for Drug Screening *In Silico*. *PLOS ONE* **6**, e15939 (2011).
- [42] Xue, R. *et al.* TCMID: traditional Chinese medicine integrative database for herb molecular mechanism analysis. *Nucleic Acids Research* **41**, D1089–D1095 (2012).
- [43] Zhang, R.-z., Yu, S.-j., Bai, H. & Ning, K. TCM-Mesh: The database and analytical system for network pharmacology analysis for TCM preparations. *Scientific Reports* **7**, 2821 (2017).
- [44] Polur, H., Joshi, T., Workman, C. T., Lavekar, G. & Kouskoumvekaki, I. Back to the Roots: Prediction of Biologically Active Natural Products from Ayurveda Traditional Medicine. *Molecular Informatics* **30**, 181–187 (2011).
- [45] Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082 (2018).

- [46] Li, B. *et al.* YaTCM: Yet another Traditional Chinese Medicine Database for Drug Discovery. *Computational and Structural Biotechnology Journal* **16**, 600–610 (2018).
- [47] Chen, Q. *et al.* SuperTCM: A biocultural database combining biological pathways and historical linguistic data of Chinese Materia Medica for drug development. *Biomedicine & Pharmacotherapy* **144**, 112315 (2021).
- [48] van Santen, J. A. *et al.* The Natural Products Atlas: An Open Access Knowledge Base for Microbial Natural Products Discovery. *ACS Central Science* **5**, 1824–1833 (2019).
- [49] van Santen, J. A. *et al.* The Natural Products Atlas 2.0: a database of microbially-derived natural products. *Nucleic Acids Research* **50**, D1317–D1323 (2022).
- [50] Mohanraj, K. *et al.* IMPPAT: A curated database of Indian Medicinal Plants, Phytochemistry And Therapeutics. *Scientific Reports* **8**, 4329 (2018).
- [51] Vivek-Ananth, R. P., Mohanraj, K., Sahoo, A. K. & Samal, A. IMPPAT 2.0: an enhanced and expanded phytochemical atlas of Indian medicinal plants. *bioRxiv* 2022.06.17.496609 (2022).
- [52] 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 Advances* **11**, 2596–2607 (2021).
- [53] 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* **25**, 3822 (2020).
- [54] Vivek-Ananth, R. P., Krishnaswamy, S. & Samal, A. Potential phytochemical inhibitors of SARS-CoV-2 helicase Nsp13: a molecular docking and dynamic simulation study. *Molecular Diversity* **26**, 429–442 (2022).

- [55] Szklarczyk, D. *et al.* STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Research* **44**, D380–D384 (2016).
- [56] 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 Advances* **12**, 6234–6247 (2022).
- [57] 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. *Scientific Reports* **10**, 17699 (2020).
- [58] Prasanth, D. S. N. B. K. *et al.* *In silico* identification of potential inhibitors from *Cinnamon* against main protease and spike glycoprotein of SARS CoV-2. *Journal of Biomolecular Structure and Dynamics* **39**, 4618–4632 (2021).
- [59] Borkotoky, S. & Banerjee, M. A computational prediction of SARS-CoV-2 structural protein inhibitors from *Azadirachta indica* (Neem). *Journal of Biomolecular Structure and Dynamics* **39**, 4111–4121 (2021).
- [60] Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016).
- [61] Macheleidt, J. *et al.* Regulation and Role of Fungal Secondary Metabolites. *Annual Review of Genetics* **50**, 371–392 (2016).
- [62] Yadav, A. N., Mishra, S., Singh, S. & Gupta, A. (eds.) *Recent advancement in white biotechnology through fungi*. Fungal biology (Springer International, 2019).
- [63] Hu, P.-F. *et al.* Oxidative Stress Induction Is a Rational Strategy to Enhance the Productivity of *Antrodia cinnamomea* Fermentations for the Antioxidant Secondary Metabolite Antrodin C. *Journal of Agricultural and Food Chemistry* **68**, 3995–4004 (2020).

- [64] Keller, N. P. Fungal secondary metabolism: regulation, function and drug discovery. *Nature Reviews Microbiology* **17**, 167–180 (2019).
- [65] Wang, F.-F. *et al.* Medicinal mushroom *Phellinus igniarius* induced cell apoptosis in gastric cancer SGC-7901 through a mitochondria-dependent pathway. *Biomedicine & Pharmacotherapy* **102**, 18–25 (2018).
- [66] Hobbs, C. *Medicinal mushrooms: an exploration of tradition, healing, and culture*. Herbs and health series (Botanica Press, Summertown, Tenn, 2003).
- [67] Powell, M. *Medicinal mushrooms: the essential guide* (Mycology Press, 2013).
- [68] Meuninck, J. *Basic illustrated edible and medicinal mushrooms* (FalconGuides, Guilford, Connecticut, 2015).
- [69] Agrawal, D. C. & Dhanasekaran, M. *Medicinal Mushrooms: Recent Progress in Research and Development* (Springer, Singapore, 2019).
- [70] Valverde, M. E., Hernández-Pérez, T. & Paredes-López, O. Edible mushrooms: improving human health and promoting quality life. *International Journal of Microbiology* **2015**, 376387 (2015).
- [71] Zaidman, B.-Z., Yassin, M., Mahajna, J. & Wasser, S. P. Medicinal mushroom modulators of molecular targets as cancer therapeutics. *Applied Microbiology and Biotechnology* **67**, 453–468 (2005).
- [72] Wishart, D. S. Introduction to cheminformatics. *Current Protocols in Bioinformatics* **18**, 14.1.1–14.1.9 (2007).
- [73] Hoffmann, T. & Gastreich, M. The next level in chemical space navigation: going far beyond enumerable compound libraries. *Drug Discovery Today* **24**, 1148–1156 (2019).
- [74] RDKit: Open-source cheminformatics. <http://www.rdkit.org>.

- [75] O’Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *Journal of Cheminformatics* **3**, 33 (2011).
- [76] Sterling, T. & Irwin, J. J. Zinc 15 – ligand discovery for everyone. *Journal of Chemical Information and Modeling* **55**, 2324–2337 (2015).
- [77] Chambers, J. *et al.* UniChem: a unified chemical structure cross-referencing and identifier tracking system. *Journal of Cheminformatics* **5**, 3 (2013).
- [78] Bemis, G. W. & Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *Journal of Medicinal Chemistry* **39**, 2887–2893 (1996).
- [79] Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of Cheminformatics* **8**, 61 (2016).
- [80] Kim, H. W. *et al.* NPClassifier: A Deep Neural Network-Based Structural Classification Tool for Natural Products. *Journal of Natural Products* **84**, 2795–2807 (2021).
- [81] NP-Likeness score.
- [82] Ertl, P., Roggo, S. & Schuffenhauer, A. Natural product-likeness score and its application for prioritization of compound libraries. *Journal of Chemical Information and Modeling* **48**, 68–74 (2008).
- [83] T. T. Tanimoto. An Elementary Mathematical theory of Classification and Prediction. *IBM* (1957).
- [84] Clemons, P. A. *et al.* Small molecules of different origins have distinct distributions of structural complexity that correlate with protein-binding profiles. *Proceedings of the National Academy of Sciences* **107**, 18787–18792 (2010).
- [85] Hoffmann, M. *et al.* SARS-CoV-2 Cell Entry Depends on ACE2 and TMPRSS2 and Is Blocked by a Clinically Proven Protease Inhibitor. *Cell* **181**, 271–280.e8 (2020).

- [86] Ou, X. *et al.* Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nature Communications* **11**, 1620 (2020).
- [87] Habtemariam, S. *et al.* Should We Try SARS-CoV-2 Helicase Inhibitors for COVID-19 Therapy? *Archives of Medical Research* **51**, 733–735 (2020).
- [88] Gurib-Fakim, A. Medicinal plants: Traditions of yesterday and drugs of tomorrow. *Molecular Aspects of Medicine* **27**, 1–93 (2006).
- [89] Petrovska, B. Historical review of medicinal plants' usage. *Pharmacognosy Reviews* **6**, 1 (2012).
- [90] Prasanth, D. S. N. B. K. *et al.* *In silico* identification of potential inhibitors from *Cinnamon* against main protease and spike glycoprotein of SARS CoV-2. *Journal of Biomolecular Structure and Dynamics* **39**, 4618–4632 (2021).
- [91] Borkotoky, S. & Banerjee, M. A computational prediction of SARS-CoV-2 structural protein inhibitors from *Azadirachta indica* (Neem). *Journal of Biomolecular Structure and Dynamics* **39**, 4111–4121 (2021).
- [92] Kalwij, J. M. Review of 'The Plant List, a working list of all plant species'. *Journal of Vegetation Science* **23**, 998–1002 (2012).
- [93] Khare, C. P. *Indian medicinal plants: an illustrated dictionary* (Springer, New York, 2007).
- [94] Kirtikar, K. R. & B., B., D. *Indian Medicinal Plants.*, vol. 1 (Periodical Experts Book Agency, 2012), 2nd edn.
- [95] Kirtikar, K. R. & B., B., D. *Indian Medicinal Plants.*, vol. 2 (Periodical Experts Book Agency, 2012), 2nd edn.
- [96] Kirtikar, K. R. & B., B., D. *Indian Medicinal Plants.*, vol. 3 (Periodical Experts Book Agency, 2012), 2nd edn.

- [97] Kirtikar, K. R. & B., B., D. *Indian Medicinal Plants.*, vol. 4 (Periodical Experts Book Agency, 2012), 2nd edn.
- [98] Kirtikar, K. R. & B., B., D. *Indian Medicinal Plants.*, vol. 5 (Periodical Experts Book Agency, 2012), 2nd edn.
- [99] Kirtikar, K. R. & B., B., D. *Indian Medicinal Plants.*, vol. 6 (Periodical Experts Book Agency, 2012), 2nd edn.
- [100] Kirtikar, K. R. & B., B., D. *Indian Medicinal Plants.*, vol. 7 (Periodical Experts Book Agency, 2012), 2nd edn.
- [101] Kirtikar, K. R. & B., B., D. *Indian Medicinal Plants.*, vol. 8 (Periodical Experts Book Agency, 2012), 2nd edn.
- [102] Duke, J. A. *Handbook of phytochemical constituents of GRAS herbs and other economic plants* (CRC Press, Boca Raton, 1992).
- [103] NCBI Resource Coordinators. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **45**, D12–D17 (2017).
- [104] Bird, S., Klein, E. & Loper, E. *Natural language processing with Python* (O'Reilly, Beijing ; Cambridge [Mass.], 2009), 1st ed edn.
- [105] Kim, S. *et al.* PubChem Substance and Compound databases. *Nucleic Acids Research* **44**, D1202–D1213 (2016).
- [106] Hastings, J. *et al.* The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research* **41**, D456–D463 (2012).
- [107] Pence, H. E. & Williams, A. Chemspider: An online chemical information resource. *Journal of Chemical Education* **87**, 1123–1124 (2010).

- [108] Nakamura, K. *et al.* KNApSAcK-3D: A Three-Dimensional Structure Database of Plant Metabolites. *Plant and Cell Physiology* **54**, e4–e4 (2013).
- [109] Linstrom, P. NIST Chemistry WebBook, NIST Standard Reference Database 69 (1997).
- [110] Wishart, D. S. *et al.* HMDB 3.0—The Human Metabolome Database in 2013. *Nucleic Acids Research* **41**, D801–D807 (2012).
- [111] Deorani, S. C. & Sharma, G. D. *Medicinal plants of Nagaland* (Bishen Singh Mahendra Pal Singh, Dehra Dun, 2007).
- [112] Hora, S. L. & Hamilton, F. *Wild Medicinal Plants of India*. (Bishen Singh Mahendra Pal Singh, Dehra Dun, 2005).
- [113] Singh, V. K. & Govil, J. N. *Recent progress in medicinal plants-Ethnomedicine and pharmacognosy.*, vol. 1 (Studium Press LLC, U.S.A., 2002), - edn.
- [114] Singh, V. K. & Govil, J. N. *Recent progress in medicinal plants-Ethnomedicine and pharmacognosy II.*, vol. 7 (Studium Press LLC, U.S.A., 2003), - edn.
- [115] Singh, V. K. & Govil, J. N. *Recent progress in medicinal plants-Ethnomedicine and pharmacognosy IV.*, vol. 22 (Studium Press LLC, U.S.A., 2008), - edn.
- [116] Singh, V. K. & Govil, J. N. *Recent progress in medicinal plants- Phytopharmacology and Therapeutric values II.* (Studium Press LLC, U.S.A., 2008).
- [117] Gupta, A. K., Tandon, N. & Sharma, M. *Quality Standards of Indian Medicinal Plants*. (Indian Council of Medical Research, 2006).
- [118] Kaushik, P. & Dhiman, A. K. *Medicinal plants and raw drugs of India* (Bishen Singh Mahendra Pal Singh, Dehra Dun, 2000).
- [119] Kshirsagar, R. D. & Singh, N. P. *Ethnobotany of Mysore and Coorg, Karnataka State* (Bishen Singh Mahendra Pal Singh, Dehradun, 2007).

- [120] Kala, C. P. *Medical plants of Indian trans-Himalaya: focus on Tibetan use of medicinal resources* (Bishen Singh Mahendra Pal Singh, Dehra Dun, 2003).
- [121] Pande, P. C., Tiwari, L. & Pande, H. C. *Folk-medicine and aromatic plants of Uttaranchal* (Bishen Singh Mahendra Pal Singh, Dehra Dun, 2006).
- [122] Sharma, U. K. *Medicinal plants of Assam* (Bishen Singh Mahendra Pal Singh, Dehra Dun, India, 2004).
- [123] Singh, K. K. & Kushal Kumar. *Ethnobotanical wisdom of Gaddi tribe in Western Himalaya* (Bishen Singh Mahendra Pal Singh, Dehradun, 2000).
- [124] Viswanathan, M. B., Prem Kumar, E. H. & Ramesh, N. *Ethnobotany of the Kannis: Kalakkad-Mundanthurai Tiger Reserve in Tirunelveli district, Tamilnadu, India* (Bishen Singh Mahendra Pal Singh, Dehra Dun, 2006).
- [125] Schriml, L. M. *et al.* Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research* **40**, D940–D946 (2012).
- [126] Hamosh, A. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research* **33**, D514–D517 (2004).
- [127] Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research* **32**, 267D–270 (2004).
- [128] Rogers, F. B. Medical subject headings. *Bulletin of the Medical Library Association* **51**, 114–116 (1963).
- [129] Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Research* **45**, D833–D839 (2017).
- [130] Zhou, X., Menche, J., Barabási, A.-L. & Sharma, A. Human symptoms–disease network. *Nature Communications* **5**, 4212 (2014).

- [131] IUCN. The IUCN Red List of Threatened Species. version 2021-3 (2021).
- [132] Kuete, V., Viertel, K. & Efferth, T. 18 - Antiproliferative Potential of African Medicinal Plants. In Kuete, V. (ed.) *Medicinal Plant Research in Africa*, 711–724 (Elsevier, Oxford, 2013).
- [133] Kinghorn, A. D. Reviews on Indian Medicinal Plants, Vols. 1-3 (Abe-Alle; Alliard; Are-Azi) Edited by A. K. Gupta and N. Tandon, Assisted by M. Sharma (Indian Council of Medical Research, New Delhi). 2004. *Journal of Natural Products* **68**, 153–154 (2005).
- [134] Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Research* **49**, D1388–D1395 (2021).
- [135] Daina, A., Michielin, O. & Zoete, V. SwissADME: a free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific Reports* **7**, 42717 (2017).
- [136] Lipkus, A. H. *et al.* Structural Diversity of Organic Chemistry. A Scaffold Analysis of the CAS Registry. *The Journal of Organic Chemistry* **73**, 4443–4451 (2008).
- [137] 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. *The Journal of Organic Chemistry* **84**, 13948–13956 (2019).
- [138] Ertl, P. An algorithm to identify functional groups in organic molecules. *Journal of Cheminformatics* **9**, 36 (2017).
- [139] O’Boyle, N. & Dalke, A. DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures. *ChemRxiv* 7097960.v1 (2018).
- [140] Sorokina, M. & Steinbeck, C. NaPLoS: a natural products likeness scorer—web application and database. *Journal of Cheminformatics* **11**, 55 (2019).

- [141] Vanii Jayaseelan, K., Moreno, P., Truszkowski, A., Ertl, P. & Steinbeck, C. Natural product-likeness score revisited: an open-source, open-data implementation. *BMC Bioinformatics* **13**, 106 (2012).
- [142] Yap, C. W. PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. *Journal of Computational Chemistry* **32**, 1466–1474 (2011).
- [143] Tweedie, S. *et al.* Genenames.org: the HGNC and VGNC resources in 2021. *Nucleic Acids Research* **49**, D939–D946 (2021).
- [144] Méndez-Lucio, O. & Medina-Franco, J. L. The many roles of molecular complexity in drug discovery. *Drug Discovery Today* **22**, 120–126 (2017).
- [145] Clemons, P. A. *et al.* Quantifying structure and performance diversity for sets of small molecules comprising small-molecule screening collections. *Proceedings of the National Academy of Sciences* **108**, 6817–6822 (2011).
- [146] Brown, N. & Jacoby, E. On scaffolds and hopping in medicinal chemistry. *Mini Reviews in Medicinal Chemistry* **6**, 1217–1229 (2006).
- [147] Zeng, X. *et al.* CMAUP: a database of collective molecular activities of useful plants. *Nucleic Acids Research* **47**, D1118–D1127 (2019).
- [148] Sorokina, M., Merseburger, P., Rajan, K., Yirik, M. A. & Steinbeck, C. COCONUT online: Collection of Open Natural Products database. *Journal of Cheminformatics* **13**, 2 (2021).
- [149] Ntie-Kang, F. *et al.* AfroDb: A Select Highly Potent and Diverse Natural Product Library from African Medicinal Plants. *PLOS ONE* **8**, e78085 (2013).
- [150] Banerjee, P. *et al.* Super Natural II—a database of natural products. *Nucleic Acids Research* **43**, D935–939 (2015).

- [151] Medina-Franco, J. L., Martínez-Mayorga, K., Bender, A. & Scior, T. Scaffold Diversity Analysis of Compound Data Sets Using an Entropy-Based Measure. *QSAR & Combinatorial Science* **28**, 1551–1560 (2009).
- [152] González-Medina, M. *et al.* Scaffold Diversity of Fungal Metabolites. *Frontiers in Pharmacology* **8** (2017).
- [153] Krier, M., Bret, G. & Rognan, D. Assessing the Scaffold Diversity of Screening Libraries. *Journal of Chemical Information and Modeling* **46**, 512–524 (2006).
- [154] Ertl, P. & Rohde, B. The Molecule Cloud - compact visualization of large collections of molecules. *Journal of Cheminformatics* **4**, 12 (2012).
- [155] Scopy. <https://scopy.iamkotori.com/>.
- [156] 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. *Advanced Drug Delivery Reviews* **46**, 3–26 (2001).
- [157] 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. *Journal of Combinatorial Chemistry* **1**, 55–68 (1999).
- [158] Veber, D. F. *et al.* Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *Journal of Medicinal Chemistry* **45**, 2615–2623 (2002).
- [159] Egan, W. J., Merz, K. M. & Baldwin, J. J. Prediction of Drug Absorption Using Multivariate Statistics. *Journal of Medicinal Chemistry* **43**, 3867–3877 (2000).
- [160] Hughes, J. D. *et al.* Physiochemical drug properties associated with in vivo toxicological outcomes. *Bioorganic & Medicinal Chemistry Letters* **18**, 4872–4875 (2008).

- [161] Gleeson, M. P. Generation of a Set of Simple, Interpretable ADMET Rules of Thumb. *Journal of Medicinal Chemistry* **51**, 817–834 (2008).
- [162] Conway, J. R., Lex, A. & Gehlenborg, N. UpSetR: an R package for the visualization of intersecting sets and their properties. *Bioinformatics* **33**, 2938–2940 (2017).
- [163] Bickerton, G. R., Paolini, G. V., Besnard, J., Muresan, S. & Hopkins, A. L. Quantifying the chemical beauty of drugs. *Nature Chemistry* **4**, 90–98 (2012).
- [164] Rogers, D. & Hahn, M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* **50**, 742–754 (2010).
- [165] Shannon, P. *et al.* Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks. *Genome Research* **13**, 2498–2504 (2003).
- [166] Schmidt, R. *et al.* Comparing Molecular Patterns Using the Example of SMARTS: Theory and Algorithms. *Journal of Chemical Information and Modeling* **59**, 2560–2571 (2019).
- [167] Ehmki, E. S. R., Schmidt, R., Ohm, F. & Rarey, M. Comparing Molecular Patterns Using the Example of SMARTS: Applications and Filter Collection Analysis. *Journal of Chemical Information and Modeling* **59**, 2572–2586 (2019).
- [168] Sayers, E. W. *et al.* Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research* **48**, D9–D16 (2020).
- [169] Roskov, Y. *et al.* Species 2000 & ITIS Catalogue of Life, 2019 Annual Checklist (2019).
- [170] Robert, V. *et al.* MycoBank gearing up for new horizons. *IMA Fungus* **4**, 371–379 (2013).
- [171] Lawrey, J. D. & Diederich, P. Lichenicolous Fungi: Interactions, Evolution, and Biodiversity. *The Bryologist* **106**, 80–120 (2003).

- [172] Merillon, J.-M. & Ramawat, K. G. *Fungal metabolites* (Springer International Publishing, Switzerland, 2017).
- [173] Pence, H. E. & Williams, A. ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education* **87**, 1123–1124 (2010).
- [174] Riniker, S. & Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *Journal of Chemical Information and Modeling* **55**, 2562–2574 (2015).
- [175] Halgren, T. A. Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. *Journal of Computational Chemistry* **17**, 490–519 (1996).
- [176] 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 1PII of original article: S0169-409X(96)00423-1. The article was originally published in *Advanced Drug Delivery Reviews* 23 (1997) 3–25. 1. *Advanced Drug Delivery Reviews* **46**, 3–26 (2001).
- [177] Teague, S. J., Davis, A. M., Leeson, P. D. & Oprea, T. The Design of Leadlike Combinatorial Libraries. *Angewandte Chemie International Edition* **38**, 3743–3748 (1999).
- [178] Grigoriev, I. V. *et al.* The Genome Portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Research* **40**, D26–D32 (2012).
- [179] Medicode (Firm) (ed.) *ICD-9-CM: International classification of diseases, 9th revision, clinical modification* (Medicode, Salt Lake City, Utah, 1997), 5th ed edn.
- [180] Bruford, E. A. *et al.* The HGNC Database in 2008: a resource for the human genome. *Nucleic Acids Research* **36**, D445–D448 (2007).

- [181] Koehn, F. E. Biosynthetic medicinal chemistry of natural product drugs. *MedChemComm* **3**, 854–865 (2012).
- [182] Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics* **7**, 20 (2015).
- [183] Jasial, S., Hu, Y., Vogt, M. & Bajorath, J. Activity-relevant similarity values for fingerprints and implications for similarity searching. *F1000Research* **5** (2016).
- [184] Cao, Y., Jiang, T. & Girke, T. A maximum common substructure-based algorithm for searching and predicting drug-like compounds. *Bioinformatics* **24**, i366–i374 (2008).
- [185] Schomburg, K., Ehrlich, H.-C., Stierand, K. & Rarey, M. From Structure Diagrams to Visual Chemical Patterns. *Journal of Chemical Information and Modeling* **50**, 1529–1535 (2010).
- [186] Huang, C. *et al.* Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The Lancet* **395**, 497–506 (2020).
- [187] Wang, C., Horby, P. W., Hayden, F. G. & Gao, G. F. A novel coronavirus outbreak of global health concern. *The Lancet* **395**, 470–473 (2020).
- [188] Zhu, N. *et al.* A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine* (2020).
- [189] Chan, J. F.-W. *et al.* A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *The Lancet* **395**, 514–523 (2020).
- [190] de Wit, E., van Doremalen, N., Falzarano, D. & Munster, V. J. SARS and MERS: recent insights into emerging coronaviruses. *Nature Reviews Microbiology* **14**, 523–534 (2016).

- [191] Kuzikov, M. *et al.* Identification of Inhibitors of SARS-CoV-2 3CL-Pro Enzymatic Activity Using a Small Molecule in Vitro Repurposing Screen. *ACS Pharmacology & Translational Science* **4**, 1096–1110 (2021).
- [192] Braga, L. *et al.* Drugs that inhibit TMEM16 proteins block SARS-CoV-2 spike-induced syncytia. *Nature* **594**, 88–93 (2021).
- [193] US Food and Drug Administration (2020) FDA Issues emergency use authorization for potential COVID-19 treatment.
- [194] Li, G. & De Clercq, E. Therapeutic options for the 2019 novel coronavirus (2019-nCoV). *Nature Reviews Drug Discovery* **19**, 149–150 (2020).
- [195] Wu, A. *et al.* Genome Composition and Divergence of the Novel Coronavirus (2019-nCoV) Originating in China. *Cell Host & Microbe* **27**, 325–328 (2020).
- [196] Wu, C. *et al.* Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods. *Acta Pharmaceutica Sinica B* **10**, 766–788 (2020).
- [197] Elfiky, A. A. Ribavirin, Remdesivir, Sofosbuvir, Galidesivir, and Tenofovir against SARS-CoV-2 RNA dependent RNA polymerase (RdRp): A molecular docking study. *Life Sciences* **253**, 117592 (2020).
- [198] Islam, R. *et al.* A molecular modeling approach to identify effective antiviral phytochemicals against the main protease of SARS-CoV-2. *Journal of Biomolecular Structure and Dynamics* **39**, 3213–3224 (2021).
- [199] Rahman, N. *et al.* Virtual Screening of Natural Products against Type II Transmembrane Serine Protease (TMPRSS2), the Priming Agent of Coronavirus 2 (SARS-CoV-2). *Molecules* **25**, 2271 (2020).
- [200] Shah, B., Modi, P. & Sagar, S. R. In silico studies on therapeutic agents for COVID-19: Drug repurposing approach. *Life Sciences* **252**, 117652 (2020).

- [201] Kumar, A. *et al.* Identification of phytochemical inhibitors against main protease of COVID-19 using molecular modeling approaches. *Journal of Biomolecular Structure and Dynamics* **39**, 3760–3770 (2021).
- [202] Liu, S., Zheng, Q. & Wang, Z. Potential covalent drugs targeting the main protease of the SARS-CoV-2 coronavirus. *Bioinformatics* **36**, 3295–3298 (2020).
- [203] Olubiyi, O. O., Olagunju, M., Keutmann, M., Loschwitz, J. & Strodel, B. High Throughput Virtual Screening to Discover Inhibitors of the Main Protease of the Coronavirus SARS-CoV-2. *Molecules* **25**, 3193 (2020).
- [204] Ahmad, M. *et al.* Prediction of Small Molecule Inhibitors Targeting the Severe Acute Respiratory Syndrome Coronavirus-2 RNA-dependent RNA Polymerase. *ACS Omega* **5**, 18356–18366 (2020).
- [205] Ren, J.-l., Zhang, A.-H. & Wang, X.-J. Traditional Chinese medicine for COVID-19 treatment. *Pharmacological Research* **155**, 104743 (2020).
- [206] Vellingiri, B. *et al.* COVID-19: A promising cure for the global panic. *Science of The Total Environment* **725**, 138277 (2020).
- [207] Shang, J. *et al.* Cell entry mechanisms of SARS-CoV-2. *Proceedings of the National Academy of Sciences* **117**, 11727–11734 (2020).
- [208] Shang, J. *et al.* Structural basis of receptor recognition by SARS-CoV-2. *Nature* **581**, 221–224 (2020).
- [209] Yan, R. *et al.* Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. *Science* **367**, 1444–1448 (2020).
- [210] Hoffmann, M. *et al.* Nafamostat Mesylate Blocks Activation of SARS-CoV-2: New Treatment Option for COVID-19. *Antimicrobial Agents and Chemotherapy* **64**, e00754–20 (2020).

- [211] Shu, T. *et al.* SARS-Coronavirus-2 Nsp13 Possesses NTPase and RNA Helicase Activities That Can Be Inhibited by Bismuth Salts. *Virologica Sinica* **35**, 321–329 (2020).
- [212] Jia, Z. *et al.* Delicate structural coordination of the Severe Acute Respiratory Syndrome coronavirus Nsp13 upon ATP hydrolysis. *Nucleic Acids Research* **47**, 6538–6550 (2019).
- [213] White, M. A., Lin, W. & Cheng, X. Discovery of COVID-19 Inhibitors Targeting the SARS-CoV-2 Nsp13 Helicase. *The Journal of Physical Chemistry Letters* **11**, 9144–9151 (2020).
- [214] Kim, S. *et al.* PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research* **47**, D1102–D1109 (2019).
- [215] Trott, O. & Olson, A. J. Autodock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of Computational Chemistry* **31**, 455–461 (2010).
- [216] Morris, G. M. *et al.* AutoDock4 and AutoDockTools4: Automated docking with selective receptor flexibility. *Journal of Computational Chemistry* **30**, 2785–2791 (2009).
- [217] Rodrigues, J., Teixeira, J., Trellet, M. & Bonvin, A. pdb-tools: a swiss army knife for molecular structures. *F1000Research* **7** (2018).
- [218] Schmidt, T., Haas, J., Cassarino, T. G. & Schwede, T. Assessment of ligand-binding residue predictions in CASP9. *Proteins: Structure, Function, and Bioinformatics* **79**, 126–136 (2011).
- [219] Chand, A., Sahoo, D. K., Rana, A., Jena, S. & Biswal, H. S. The Prodigious Hydrogen Bonds with Sulfur and Selenium in Molecular Assemblies, Structural

- Biology, and Functional Materials. *Accounts of Chemical Research* **53**, 1580–1592 (2020).
- [220] Sarkhel, S. & Desiraju, G. R. N-H...O, O-H...O, and C-H...O hydrogen bonds in protein-ligand complexes: Strong and weak interactions in molecular recognition. *Proteins: Structure, Function, and Bioinformatics* **54**, 247–259 (2003).
- [221] Zhou, P., Tian, F., Lv, F. & Shang, Z. Geometric characteristics of hydrogen bonds involving sulfur atoms in proteins. *Proteins: Structure, Function, and Bioinformatics* **76**, 151–163 (2009).
- [222] Kříž, K., Fanfrlík, J. & Lepšík, M. Chalcogen Bonding in Protein-Ligand Complexes: PDB Survey and Quantum Mechanical Calculations. *ChemPhysChem* **19**, 2540–2548 (2018).
- [223] Borozan, S. Z. & Stojanović, S. Đ. Halogen bonding in complexes of proteins and non-natural amino acids. *Computational Biology and Chemistry* **47**, 231–239 (2013).
- [224] Freitas, R. F. d. & Schapira, M. A systematic analysis of atomic protein–ligand interactions in the PDB. *MedChemComm* **8**, 1970–1981 (2017).
- [225] Abraham, M. J. *et al.* GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19–25 (2015).
- [226] Schmid, N. *et al.* Definition and testing of the GROMOS force-field versions 54A7 and 54B7. *European Biophysics Journal* **40**, 843 (2011).
- [227] Stroet, M. *et al.* Automated Topology Builder Version 3.0: Prediction of Solvation Free Enthalpies in Water and Hexane. *Journal of Chemical Theory and Computation* **14**, 5834–5845 (2018).

- [228] Berendsen, H. J. C., Postma, J. P. M., van Gunsteren, W. F., DiNola, A. & Haak, J. R. Molecular dynamics with coupling to an external bath. *The Journal of Chemical Physics* **81**, 3684–3690 (1984).
- [229] Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: A new molecular dynamics method. *Journal of Applied Physics* **52**, 7182–7190 (1981).
- [230] Paoloni-Giacobino, A., Chen, H., Peitsch, M. C., Rossier, C. & Antonarakis, S. E. Cloning of the TMPRSS2 Gene, Which Encodes a Novel Serine Protease with Transmembrane, LDLRA, and SRCR Domains and Maps to 21q22.3. *Genomics* **44**, 309–320 (1997).
- [231] Evnin, L. B., Vásquez, J. R. & Craik, C. S. Substrate specificity of trypsin investigated by using a genetic selection. *Proceedings of the National Academy of Sciences* **87**, 6659–6663 (1990).
- [232] Waterhouse, A. *et al.* SWISS-MODEL: homology modelling of protein structures and complexes. *Nucleic Acids Research* **46**, W296–W303 (2018).
- [233] Herter, S. *et al.* Hepatocyte growth factor is a preferred *in vitro* substrate for human hepsin, a membrane-anchored serine protease implicated in prostate and ovarian cancers. *Biochemical Journal* **390**, 125–136 (2005).
- [234] Pettersen, E. F. *et al.* UCSF Chimera-A visualization system for exploratory research and analysis. *Journal of Computational Chemistry* **25**, 1605–1612 (2004).
- [235] Chen, V. B. *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography* **66**, 12–21 (2010).
- [236] Fujishima, A. *et al.* The crystal structure of human cathepsin L complexed with E-64. *FEBS Letters* **407**, 47–50 (1997).

- [237] Turk, V. *et al.* Cysteine cathepsins: From structure, function and regulation to new frontiers. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics* **1824**, 68–88 (2012).
- [238] Adams-Cioaba, M. A., Krupa, J. C., Xu, C., Mort, J. S. & Min, J. Structural basis for the recognition and cleavage of histone H3 by cathepsin L. *Nature Communications* **2**, 197 (2011).
- [239] Schyman, P., Liu, R., Desai, V. & Wallqvist, A. vNN Web Server for ADMET Predictions. *Frontiers in Pharmacology* **8** (2017).
- [240] Tsai, Y.-C. *et al.* Antiviral Action of Tryptanthrin Isolated from *Strobilanthes cusia* Leaf against Human Coronavirus NL63. *Biomolecules* **10**, 366 (2020).
- [241] Zhao, D.-G. *et al.* Coumarins with α -glucosidase and α -amylase inhibitory activities from the flower of *Edgeworthia gardneri*. *Fitoterapia* **107**, 122–127 (2015).
- [242] Gao, D. *et al.* In vitro evaluation of dual agonists for PPAR γ / β from the flower of *Edgeworthia gardneri* (wall.) Meisn. *Journal of Ethnopharmacology* **162**, 14–19 (2015).
- [243] Gao, D. *et al.* The flower of *Edgeworthia gardneri* (wall.) Meisn. suppresses adipogenesis through modulation of the AMPK pathway in 3T3-L1 adipocytes. *Journal of Ethnopharmacology* **191**, 379–386 (2016).
- [244] Gupta, P. C., Sharma, N. & Rao, C. V. A review on ethnobotany, phytochemistry and pharmacology of *Fumaria indica* (Fumitory). *Asian Pacific Journal of Tropical Biomedicine* **2**, 665–669 (2012).
- [245] Kardos, J., Blaskó, G., Kerekes, P., Kovács, I. & Simonyi, M. Inhibition of [3H]GABA binding to rat brain synaptic membranes by bicuculline related alkaloids. *Biochemical Pharmacology* **33**, 3537–3545 (1984).

- [246] Mitra, S., Shukla, V. J. & Acharya, R. Effect of Shodhana (processing) on Kupeelu (Strychnos nux-vomica Linn.) with special reference to strychnine and brucine content. *AYU (An international quarterly journal of research in Ayurveda)* **32**, 402 (2011).
- [247] Ueno, S., Bracamontes, J., Zorumski, C., Weiss, D. S. & Steinbach, J. H. Bicuculline and Gabazine Are Allosteric Inhibitors of Channel Opening of the GABAA Receptor. *Journal of Neuroscience* **17**, 625–634 (1997).
- [248] Yadav, J. P. *et al.* Cassia occidentalis L.: A review on its ethnobotany, phytochemical and pharmacological profile. *Fitoterapia* **81**, 223–230 (2010).
- [249] Zhu, S. & Yang, Z. Application of Anthraquinone Derivative in Resisting Influenza Virus and Bird Flu Virus H5N1. *China Patent CN20051134677* (2005).
- [250] Mete, I. E. & Gözler, T. (+)-Oxoturkiyenine: an Isoquinoline-Derived Alkaloid from Hypecoum pendulum. *Journal of Natural Products* **51**, 272–274 (1988).
- [251] Gurung, P. & De, P. Spectrum of biological properties of cinchona alkaloids: A brief review. *Journal of Pharmacognosy and Phytochemistry* **6** (2017).
- [252] Tripathi, Y. C., Maurya, S. K., Singh, V. P. & Pandey, V. B. Cyclopeptide alkaloids from Zizyphus rugosa bark. *Phytochemistry* **28**, 1563–1565 (1989).
- [253] Yadav, A. & Singh, P. Analgesic and anti-inflammatory activities of Zizyphus rugosa root barks. *Journal of Chemical and Pharmaceutical Research* **2**, 255–259 (2010).
- [254] Wang, J.-H., Luan, F., He, X.-D., Wang, Y. & Li, M.-X. Traditional uses and pharmacological properties of Clerodendrum phytochemicals. *Journal of Traditional and Complementary Medicine* **8**, 24–38 (2018).

- [255] Sumthong, P., Romero-González, R. R. & Verpoorte, R. Identification of Anti-Wood Rot Compounds in Teak (*Tectona grandis* L.f.) Sawdust Extract. *Journal of Wood Chemistry and Technology* **28**, 247–260 (2008).
- [256] Jain, M. *et al.* Traditional uses, phytochemistry and pharmacology of *Tecomella undulata*– A review. *Asian Pacific Journal of Tropical Biomedicine* **2**, S1918–S1923 (2012).
- [257] Cadelis, M. M. *et al.* Discovery and preliminary structure–activity relationship studies on tecomaquinone I and tectol as novel farnesyltransferase and plasmodial inhibitors. *Bioorganic & Medicinal Chemistry* **24**, 3102–3107 (2016).
- [258] Bosisio, E., Benelli, C. & Pirola, O. Effect of the flavanolignans of *Silybum marianum* L. On lipid peroxidation in rat liver microsomes and freshly isolated hepatocytes. *Pharmacological Research* **25**, 147–165 (1992).
- [259] Bahmani, M., Shirzad, H., Rafeian, S. & Rafeian-Kopaei, M. *Silybum marianum* : Beyond Hepatoprotection. *Journal of Evidence-Based Complementary & Alternative Medicine* **20**, 292–301 (2015).
- [260] Wlodawer, A. *et al.* Ligand-centered assessment of SARS-CoV-2 drug target models in the Protein Data Bank. *The FEBS Journal* **287**, 3703–3718 (2020).
- [261] Daura, X. *et al.* Peptide Folding: When Simulation Meets Experiment. *Angewandte Chemie International Edition* **38**, 236–240 (1999).
- [262] *The wealth of India: a dictionary of Indian raw materials & industrial products. First supplement series (Raw materials)*, vol. 1 (National Institute of Science Communication, Council of Scientific & Industrial Research, New Delhi, 2000).
- [263] Mohd Jamil, M. D. H., Taher, M., Susanti, D., Rahman, M. A. & Zakaria, Z. A. Phytochemistry, Traditional Use and Pharmacological Activity of *Picrasma quassioides*: A Critical Reviews. *Nutrients* **12**, 2584 (2020).

- [264] Chen, J. *et al.* Tobacco Mosaic Virus (TMV) Inhibitors from *Picrasma quassioides* Benn. *Journal of Agricultural and Food Chemistry* **57**, 6590–6595 (2009).
- [265] Hoang, V. D. *et al.* Natural anti-HIV agents—part I: (+)-demethoxyepiexcelsin and verticillatol from *Litsea verticillata*. *Phytochemistry* **59**, 325–329 (2002).
- [266] Guan, Y. *et al.* Litsea Species as Potential Antiviral Plant Sources. *The American Journal of Chinese Medicine* **44**, 275–290 (2016).
- [267] Dutta, P. K., Banerjee, D. & Dutta, N. L. Euphorbetin: a new bicoumarin from *Euphorbia lathyris* L. *Tetrahedron Letters* **13**, 601–604 (1972).
- [268] Rastogi, R. P. & Mehrotra, B. N. *Compendium of Indian Medicinal Plants. Volume 3: 1980 - 1984* (Central Drug Research Institute, Lucknow, 1993).
- [269] *The wealth of India: a dictionary of Indian raw materials & industrial products. First supplement series (Raw materials)*, vol. 2 (National Institute of Science Communication, Council of Scientific & Industrial Research, New Delhi, 2000).
- [270] Yuan, H., Ma, Q., Ye, L. & Piao, G. The Traditional Medicine and Modern Medicine from Natural Products. *Molecules* **21**, 559 (2016).
- [271] Newman, D. J. & Cragg, G. M. Natural Products As Sources of New Drugs over the 30 Years from 1981 to 2010. *Journal of Natural Products* **75**, 311–335 (2012).
- [272] Moumbock, A. F., Li, J., Mishra, P., Gao, M. & Günther, S. Current computational methods for predicting protein interactions of natural products. *Computational and Structural Biotechnology Journal* **17**, 1367–1376 (2019).
- [273] Bagherian, M. *et al.* Machine learning approaches and databases for prediction of drug–target interaction: a survey paper. *Briefings in Bioinformatics* **22**, 247–269 (2020).

- [274] Nantasenamat, C., Isarankura-Na-Ayudhya, C. & Prachayasittikul, V. Advances in computational methods to predict the biological activity of compounds. *Expert Opinion on Drug Discovery* **5**, 633–654 (2010).
- [275] Stumpfe, D., Hu, H. & Bajorath, J. Evolving concept of activity cliffs. *ACS Omega* **4**, 14360–14368 (2019).
- [276] Turk, V. NEW EMBO MEMBERS' REVIEW: Lysosomal cysteine proteases: facts and opportunities. *The EMBO Journal* **20**, 4629–4633 (2001).
- [277] Genheden, S. & Ryde, U. The MM/PBSA and MM/GBSA methods to estimate ligand-binding affinities. *Expert Opinion on Drug Discovery* **10**, 449–461 (2015).
- [278] Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. Electrostatics of nanosystems: Application to microtubules and the ribosome. *Proceedings of the National Academy of Sciences* **98**, 10037–10041 (2001).
- [279] Kumari, R., Kumar, R., Open Source Drug Discovery Consortium & Lynn, A. *g_mmpbsa* —A GROMACS Tool for High-Throughput MM-PBSA Calculations. *Journal of Chemical Information and Modeling* **54**, 1951–1962 (2014).