Molecular Diversity
https://doi.org/10.1007/s11030-022-10506-5

ORIGINAL ARTICLE

Towards systematic exploration of chemical space: building the fragment library module in molecular property diagnostic suite

Anamika Singh Gaur[1,2,3] · Lijo John[1,2,3] · Nandan Kumar[1] · M. Ram Vivek[2] · Selvaraman Nagamani[1,3] · Hridoy Jyoti Mahanta[1,3] · G. Narahari Sastry[1,3]

Received: 11 July 2022 / Accepted: 23 July 2022

Abstract

A fragment-based drug discovery (FBDD) approach has traditionally been of utmost significance in drug design studies. It allows the exploration of large chemical space to find novel scaffolds and chemotypes which can be improved into selective inhibitors with good affinity. In the current work, several public domain chemical libraries (ChEMBL, DrugCentral, PDB ligands, COCONUT, and SAVI) comprising bioactive and virtual molecules were retrieved to develop a fragment library. A systematic fragmentation method that breaks a given molecule into rings, linkers, and substituents was used to cleave the molecules and the fragments were analyzed. Further, only the ring framework was taken into the consideration to develop a fragment library that consists of a total number of 107,614 unique fragments. This set represents a rich diverse structure framework that covers a wide variety of yet-to-be-explored fragments for a wide range of small molecule-based applications. This fragment library is an integral part of the molecular property diagnostic suite (MPDS) suite that can be used with other modeling and informatics methods for FBDD approaches. The fragment library module of MPDS can be accessed at http://mpds.neist.res.in:8085.

✉ G. Narahari Sastry
gnsastry@gmail.com; gnsastry@neist.res.in

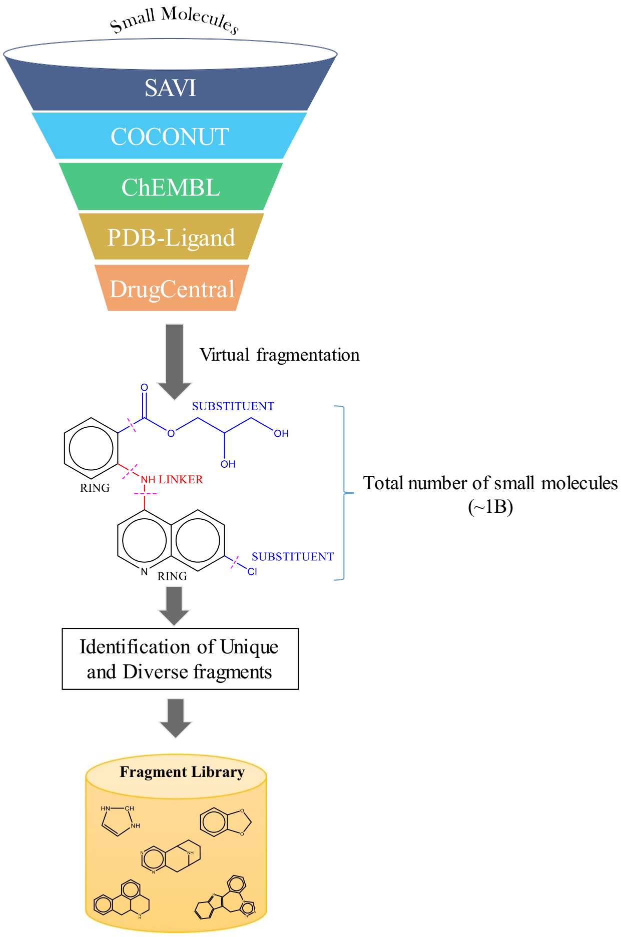
[1] Advanced Computation and Data Sciences Division,
CSIR –North East Institute of Science and Technology,
Jorhat 785006, India

[2] Centre for Molecular Modeling, CSIR-Indian Institute
of Chemical Technology, Hyderabad 500007, India

[3] Academy of Scientific and Innovative Research (AcSIR),
Ghaziabad 201002, India

Published online: 04 August 2022

Graphical abstract



Generation of MPDS fragment library from biological databases.

Keywords Fragment library · Drug discovery · Fragment space · MPDS

Introduction

Drug discovery is one of the great challenges and molecular modeling approaches that aim to identify lead compounds. A typical chemical space size is estimated to be 10^{60} molecules, and screening these molecules experimentally for the identification of lead molecules is a mammoth task. Virtual screening applies a series of filters to identify potential lead compounds from a huge pool of compounds [1–3]. Many commercial and open-source drug discovery software are available for drug discovery to minimize time and cost. Development of open-source drug discovery software along with disease-specific information, a robust compound library, and fragment library is essential for the identification of potential lead compounds with improved activity. One such effort from our group is the development of open-source computational drug discovery software MPDS [4, 5].

Fragment-based drug design (FBDD) is an effective method for quick and precise identification of chemical moieties to design selective lead molecules with high affinity [6, 7]. The outcome of FBDD is directly influenced by the composition of the fragment library being used [8, 9]. Several methods are available for the fragmentation of molecules and they are broadly categorized into four categories based on the pattern of fragmentation such as (a) hierarchical and systematic, (b) retrosynthetic, (c) knowledge-based, and (d) random fragmentation. All these four methods use the molecular connectivity for generating fragments. The systematic graph fragmentation and computer-generated fragments based on retrosynthetic rules are the two most widely used methods for constructing the fragment libraries in FBDD. The graph fragmentation cleaves the molecules at topologically defined positions such as the bond between a ring and its substituents, whereas computer-generated retrosynthesis uses defined rules based on the chemical reactions and fragments, such as amide bonds. Bemis and Murcko have examined the feature of drug molecules based on the rings, linkers, frameworks, and side-chain atoms as well as based on the atom type, hybridization, and bond order of molecules [10, 11]. They emphasized the most common frameworks that represent the structural variation in the drug molecule dataset. According to Murcko, (i) ring systems are cycles that share the edges, (ii) atoms that connect two ring systems are linkers, (iii) substituents are atoms that are neither ring nor linkers. Physiochemical properties-based rules are also employed by many people to construct fragment libraries. Congreve et al. [12], discovered the "rule of three," viz. (a) Molecular weight (MW) ≤ 300 Da; (b) Hydrogen bond donor (HBD) and Hydrogen bond acceptor (HBA) ≤ 3; (c) Log$P$ ≤ 3.42 to define the physiochemical properties of a molecule. Several molecular fragmentation techniques are being developed by researchers to obtain synthetically feasible chemical motifs/fragments (Table S1). Among these fragmentation methods, RECAP and BRICS are retrosynthetic fragmentation methods. In RECAP, molecules are fragmented based on a group of eleven defined bond categories, and make certain rings that are key structural moieties of molecules. However, BRICS rules are the expansion of the RECAP rules that consider the environment of every bond type and its surrounding substructures [13, 14]. In addition, some other fragmentation methods and unexplored fragments spaces are need to be explored [15–17]. Consequently, many fragmentation techniques for FBDD have been developed which consider binding site and fragment connection information from a macromolecule-ligand complex [18–25]. Our group has also made a series of fragment and structure-based studies using the traditional computer-aided drug design, artificial intelligence, and machine learning approaches in probing the molecular or structural properties of small molecules, macromolecules, and other complexes [26–33]. The applicability of these fragmentation methods depends on the specific purpose for which they are being used or implemented. Several approaches have been developed for the construction of fragment libraries and most of the commercial fragment libraries were found to obey widely accepted "rule of three." Over the years, various computer programs were also developed to enumerate the molecules, and currently, databases of the order of 10^{20} are created and considered ultra-large chemical repositories [34, 35]. Our approach is based on the identification of several types of fragments through a systematic fragmentation method that breaks a given molecule into rings, linkers, and substituents to understand the diversity of chemical space. We explored the vast chemical libraries that includes known and bioactive molecules namely ChEMBL [36], DrugCentral [37], PDB ligands [38], COCONUT [39], and molecules from the SAVI database [40]. Most of these fragmentation tools are not publicly available as web services and thus cannot be used by many computational drugs design. Elead3D and ACFIS, screen inbuilt fragment libraries against targets for FBDD are the few available computational drug design web services [41, 42]. While virtual screening facilities are available for fragments, there is a lack of a web server that offers fragment library construction tools, inbuilt fragment libraries, and screening facilities all together in a single platform. This impelled us to develop and integrate the FBDD module in the molecular property diagnostic suite (MPDS) suite of web portal [3–5].

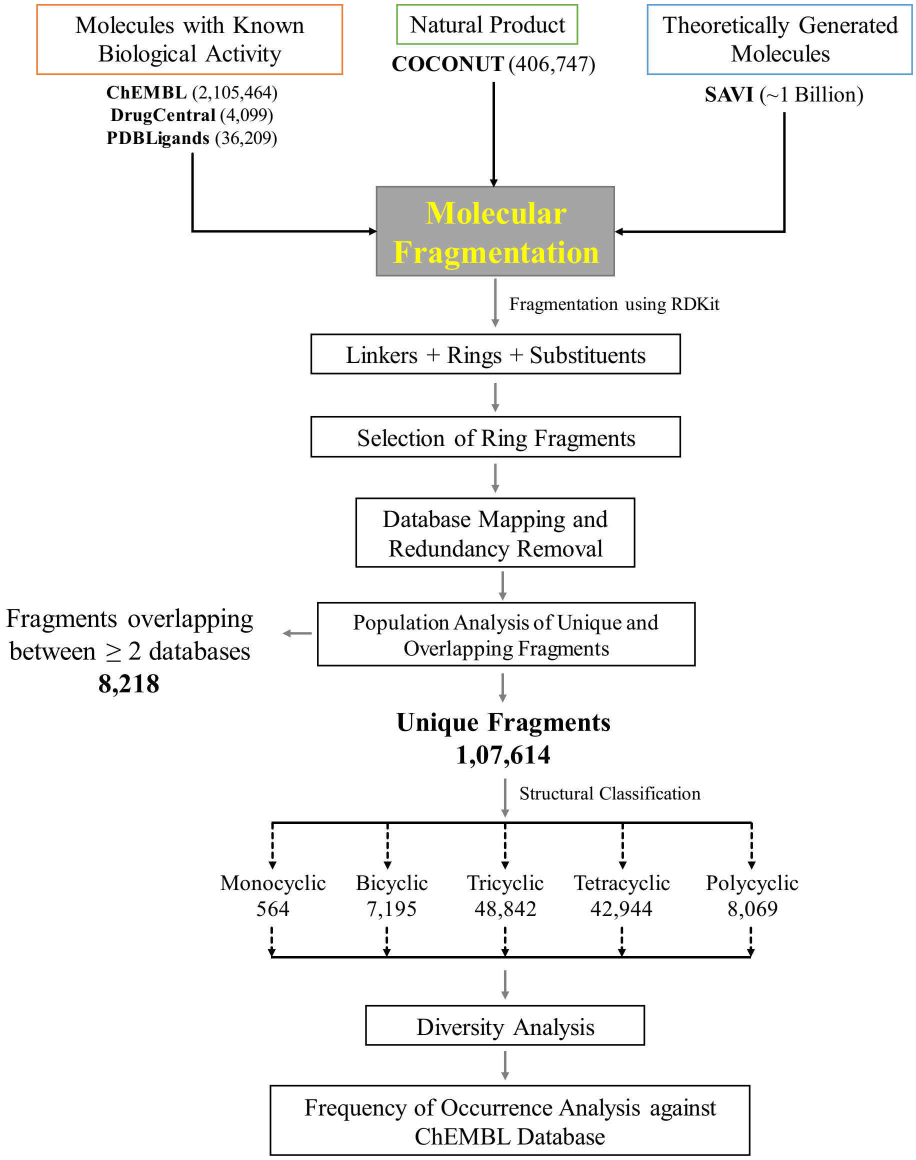
Materials and methods

Data curation and fragment generation

The bioactive molecules from ChEMBL (2,105,464), DrugCentral (4099), and PDB ligands (36,209), a set of natural products from COCONUT (406,747), and theoretically generated molecules from the SAVI database (~ 1 Billion) were retrieved. The database identifiers for all these molecules were retained and molecules from each of these databases were individually subjected to fragmentation for obtaining a set of rings, linkers, and substituents using an in-house python script that uses the "*fragmentonbonds*" function in RDkit. As the ring systems obtained from the fragmentation represent rigid entities that create the scaffold of a compound, all further analyses focused only on rings. The fragmentation algorithm in RDkit cleaves the bond between the atoms that are part of a ring and atoms that are not a part of the ring which results in the generation of small fragments [43]. We have filtered out the fragments based on the rule of three [12] thus reducing the number to a limited number for detailed analyses. All the ring fragments were further classified into structural categories, such as monocyclic, bicyclic, tricyclic, tetracyclic, and polycyclic. The complete procedure used in this study is depicted in Scheme 1.

De-duplication of ring fragments and fingerprint generation

After fragmentation, the SMILES of the ring fragments contained the patterns like "[*1]" or "[*10]" that indicated the position where the bond was cleaved. Hence, these SMILES were first cleaned to remove all such patterns, and thereafter it was subjected to the computation of InChIKey using OpenBabel3 [44]. A two-step redundancy removal was carried out, first at the database level, and second in the structural category groups using an in-house python script.


Molecules with Known Biological Activity
ChEMBL (2,105,464)
DrugCentral (4,099)
PDBLigands (36,209)
Natural Product
COCONUT (406,747)
Theoretically Generated Molecules
SAVI (~1 Billion)
Molecular Fragmentation
Fragmentation using RDKit
Linkers + Rings + Substituents
Selection of Ring Fragments
Database Mapping and Redundancy Removal
Population Analysis of Unique and Overlapping Fragments
Fragments overlapping between ≥ 2 databases
8,218
Unique Fragments
1,07,614
Structural Classification
Monocyclic
564
Bicyclic
7,195
Tricyclic
48,842
Tetracyclic
42,944
Polycyclic
8,069
Diversity Analysis
Frequency of Occurrence Analysis against ChEMBL Database

Scheme 1 The integrative methodology used for the generation of the fragment library

As InChIKey is considered to be unique for a molecule, it was aptly used to de-duplicate all the redundant ring fragments while the database ids of all the duplicate rings were retained. The information from the database ids was used to generate a 5-Bit fingerprint, and the value ‘1’ shows the availability and ‘0’ for the non-availability of the particular ring fragment in all the five databases.

Probing the unique and overlapping ring fragments

The diversity of ring fragments were further analyzed at both the database and structural category level using the fingerprints. At the database level, rings that were common or found in 2 or more databases, and rings that were unique to a specific database were identified. While at the structural level, the biologically active and ring systems generated through de novo approaches were analyzed. The Tanimoto coefficient scores were calculated using DataWarrior tool [45] and the scores were extensively used to fetch the top diverse fragments as well as their frequency of occurrence.

Similarity analysis

The unique set of rings that were identified after de-duplication were also analyzed on the basis of the Morgan fingerprints. Tanimoto coefficient was computed to generate a matrix, which was further analyzed for ranking the rings based on similarity scores.

Results and discussion

Generation of the molecular fragments

The molecular dataset were retrieved by compiling several public domain databases such as ChEMBL, DrugCentral, PDB ligands, COCONUT, and SAVI. Further, using an in-house python script these molecules were fragmented into a set of rings, linkers, and substituents. The cyclic fragments (interconnected ring systems) were classified into a set of monocyclic, bicyclic, tricyclic, tetracyclic, and polycyclic categories which was necessary for introspecting the structural diversity of fragments. A total of 107,614 non-redundant ring fragments were obtained and based on the database identifier, a 5-Bit fingerprint was generated. These bit positions indicate each database, starting from ChEMBL, followed by COCONUT, DrugCentral, PDB Ligands, and SAVI. An example of a fragmentation procedure is shown in Fig. 1.

Analysis using the fingerprint

The 5-bit fingerprint was used to map both the unique and common set of molecules that overlapped among 2 or more databases. We were able to observe which type of ring fragment was very unique and which fragments were spread across the databases through this analysis. Figure 2 shows the pie-plot showing the percentage of overlap of fragments among the databases, and this indicates that the maximum overlap is from the tricyclic (62.3%), followed by bicyclic (20.1%), tetracyclic (10.70%), monocyclic (4.3%), and polycyclic (2.5%) groups. The number of overlapped fragments between ≥ 2 databases is listed in Table S2. A large overlap was observed between ChEMBL and COCONUT database, with a total of 45.53% compounds, followed by the overlap between ChEMBL and SAVI, and so on (Table S2). It is also noted that the rings from DrugCentral and PDB-Ligand datasets have least or no overlap with other databases, indicating that the core scaffolds of drugs and drug-like molecules share very less structural similarity. Thus, exploring novel scaffolds from COCONUT, ChEMBL, and SAVI, followed by detailed analysis for its targets will be a promising task to endeavor in small molecule-based drug discovery approaches.

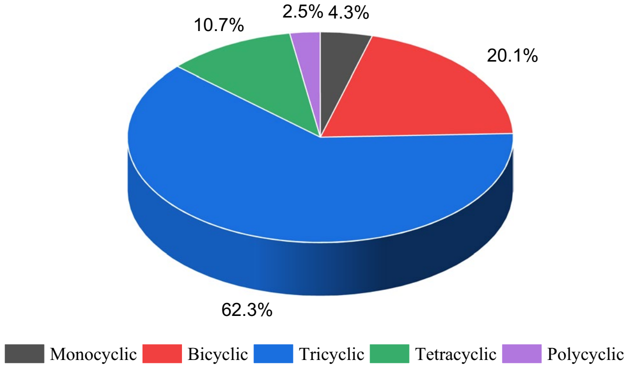


Fig. 2 The percentage of common fragments based on its structural diversity

Analysis of unique fragments

Structurally classified fragments are summarized in Table 1, Table S3, and Fig. 3. It can be observed that the maximum of the ring fragments is from the tricyclic (48,842) followed by tetracyclic (42,944), polycyclic (8069), bicyclic (7195), and monocyclic (564) groups. Subsequently, database-wise distribution of these fragments showed that the maximum

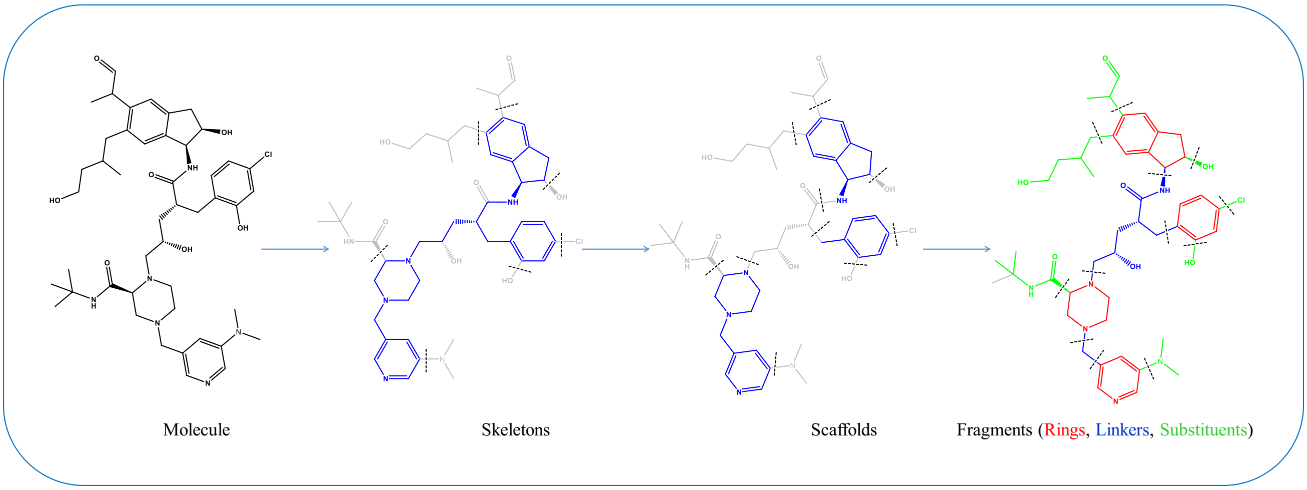


Fig. 1 Process of generation of fragments from molecules into rings, linkers, and substituents

Table 1 The population of different types of structurally unique fragments

Dataset	Monocyclic	Bicyclic	Tricyclic	Tetracyclic	Polycyclic	Total
ChEMBL	355	202	14,452	7398	1658	24,065
COCONUT	173	2156	3018	2098	1354	8799
DrugCentral	0	33	3	0	2	38
PDB Ligands	16	280	259	127	16	698
SAVI	20	4524	31,110	33,321	5039	74,014
Total	564	7195	48,842	42,944	8069	1,07,614

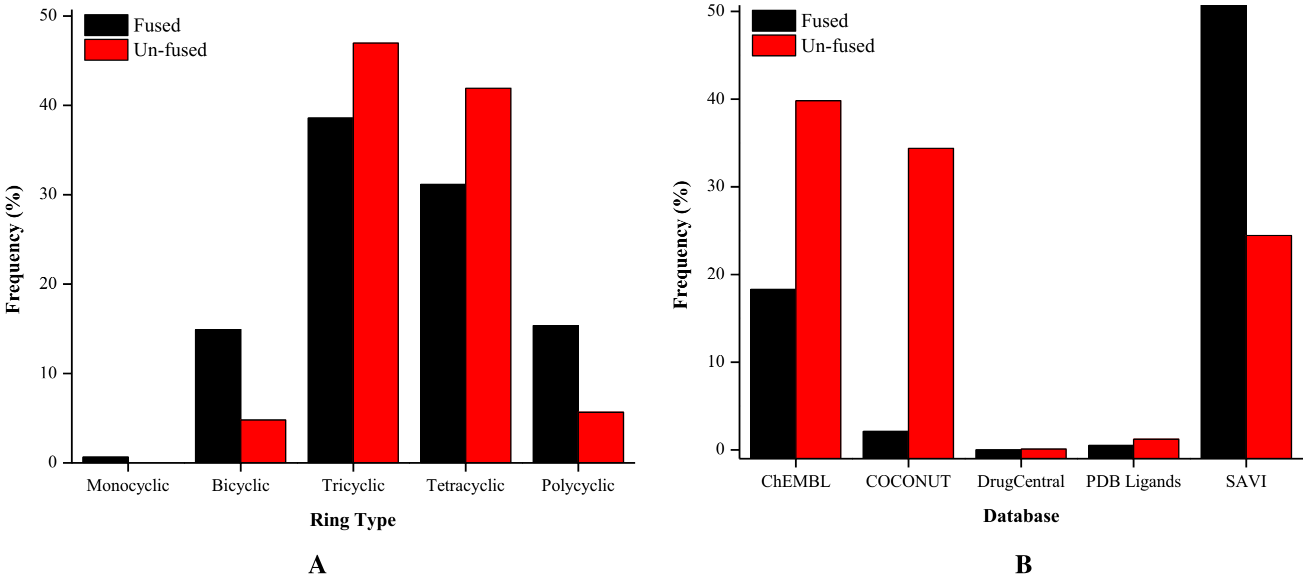


Fig. 3 Percentage distribution of the ring fragments in the dataset based on **A** type of ring and **B** different databases

number of unique fragments have been obtained from SAVI (74,014), followed by ChEMBL (24,065), COCONUT (8799), PDB ligands (698), and DrugCentral (38). This analysis suggests that SAVI, ChEMBL, COCONUT, and PDB ligands databases are very important in terms of structural diversity, and they are key resources that can be used for the development of a highly diverse fragment library.

Analysis of structural diversity and frequency of occurrence

The diversity of non-redundant fragments was investigated by comparing the fragments to each other and calculating the all-by-all similarity matrix using the Tanimoto coefficient. Analysis and interpretation of the resultant matrix (107,614 × 107,614) is difficult and therefore we have generated a separate matrix for only fused fragments. Subsequently, the data point of the matrix has been arranged in a 2D plane by grouping the distribution of Tanimoto scores, as shown in Figure S1. The results exhibited that the Tanimoto score of ~ 80% and ~ 18% of the fragments are nearly 0.1 and 0.2, respectively, i.e., ~ 98% of the fragments are highly dissimilar, suggesting that the generated library of the ring fragments (20,225) is very diverse.

It showed that tricyclic ring fragments are a more diverse dataset followed by tetracyclic, polycyclic, and bicyclic ring fragments in terms of the Tanimoto coefficient. Followed by these analyses, the diverse fragments are ranked within each structurally classified group using algorithms implemented in the DataWarrior software and the highly ranked fragments are subjected to the frequency of occurrence analysis against the ChEMBL database. In Fig. 4, the top ten most frequent unique fragments from monocyclic, bicyclic, tricyclic, tetracyclic, and polycyclic groups that are biologically active as found in the ChEMBL database are represented. The highest occurrence is observed for the monocyclic ring fragments followed by the tricyclic, bicyclic, tetracyclic, and polycyclic ring fragments. Among the top ten fragments from each of these groups, the highest occurrence is found in the monocyclic group with 0.684% (14,391) fragments, and the lowest occurrence is found in the polycyclic group with 0.003% (59) fragments. It is also observed that in all the structural categories, the 10 most frequent ring fragments are dominated by the hetero-aliphatic

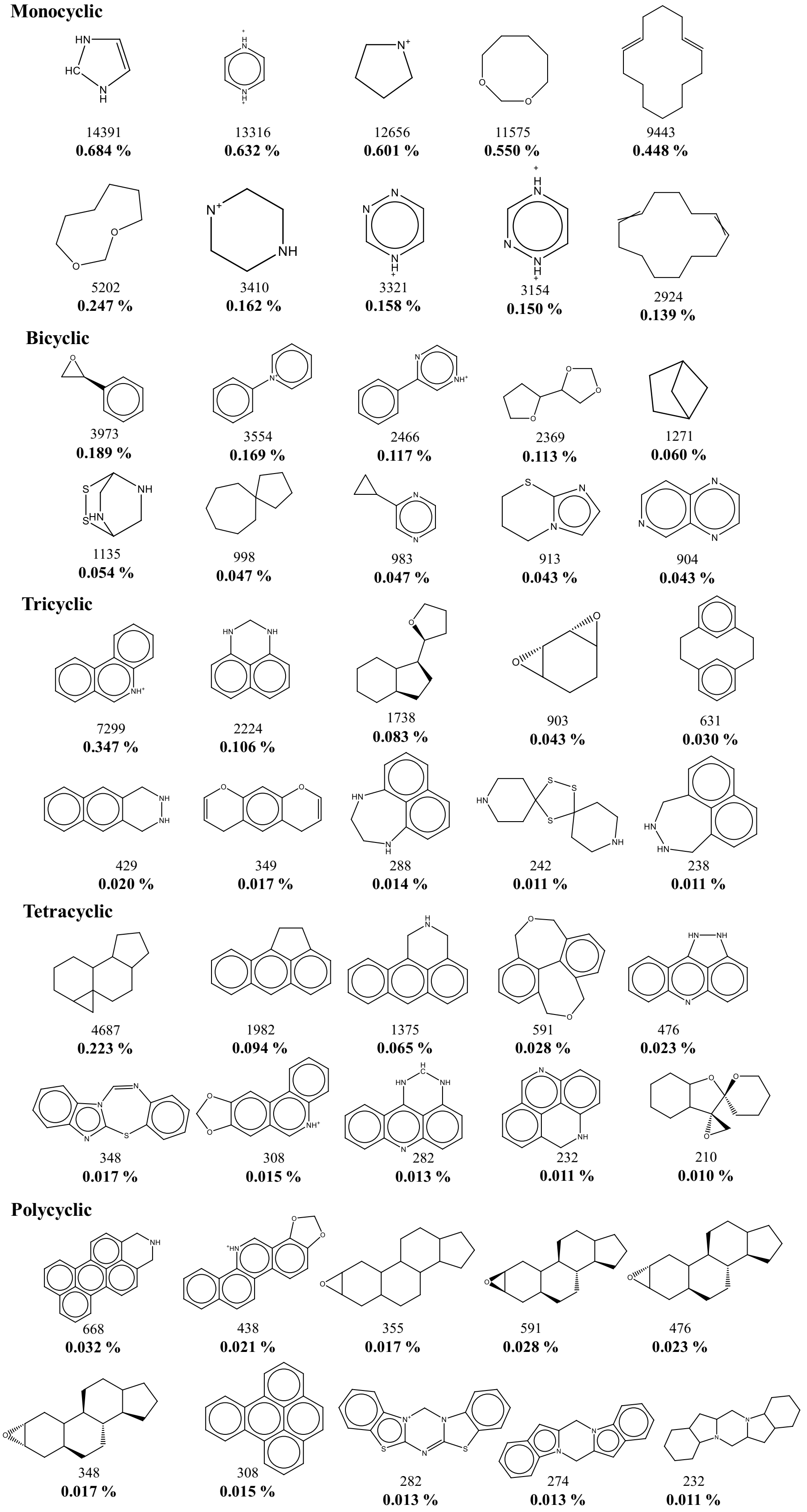


Fig. 4 Ten most frequent unique fragments from monocyclic, bicyclic, tricyclic, tetracyclic, and polycyclic groups against the ChEMBL database. The occurrence of these fragments in the ChEMBL dataset is indicated in the normal font and the percentage in bold font

rings in comparison to carbo-aliphatic rings. However, the hetero-aliphatic and carbo-aliphatic rings seemed to be highly comparable in the case of large ring size fragments. The observation is in good accordance with the observation made for GSK medicinal compounds [46]. The dominance of hetero-aliphatic rings may be preferred because of their higher hydrophilicity which often leads to improving the solubility of the chemical compounds [47], and also due to being functional in the forming of hydrogen bonds and other interactions with the biological systems. As the hetero-aliphatic rings are observed to be an important component of biologically important molecules, we have extended our analysis towards examining the count of different heteroatoms such as N, P, O, and S in the ring structure, as shown in Fig. 5 and Table S4. Results showed that the presence of N and O or the combination (N + O) are the highly occurred heteroatoms (i.e., 75.37%) in the ring fragments followed by the ring containing N in combination with S (17.58%). Interestingly, it has been observed that the distribution of carbo-cyclic ring structures is higher (0.79%) than the hetero-aliphatic rings containing S or P or O + S or the ring containing more than two heteroatoms (Fig. 5). This distribution of carbo-cyclic ring structure might be coming from the set of bigger ring size fragments as the occurrence of carbo-aliphatic rings seemed to be highly comparable to hetero-aliphatic rings in the case of bigger ring size fragments (Fig. 5). Overall, results suggested that the diversity of ring fragments is majorly offered by the hetero-aliphatic rings and the N and O containing rings are among the most significant structural component.

Incorporation of fragment library module in molecular property diagnostic suite

Molecular property diagnostic suite is a galaxy-based open-source computational drug discovery software developed in our group. The MPDS has been generated for different diseases namely tuberculosis [3, 4], diabetes [5], COVID-19, etc. The modules in MPDS are classified into disease-dependent and disease-independent modules. The disease-dependent modules are customized to specific diseases and gene information, drug information, etc., are majorly available in this module. The disease-independent modules are majorly drugged discovery modules such as molecular docking and binding site prediction. The MPDS also has a compound library of million compounds with structural classification.

The identified fragments have been incorporated into the MPDS under the MPDS fragment library section. A total of 107,614 fragments are classified into monocyclic, bicyclic, tricyclic, tetracyclic, and polycyclic. Different properties such as molecular weight, no. of hydrogen bond donors, no. hydrogen bond acceptors, molar refractivity, number of heavy atoms, number of rotatable bonds, log*P*, and topological polar surface area (TPSA) have been calculated for each fragment and incorporated into the MySQL database. The users can select types of rings (i.e., monocyclic, bicyclic, tricyclic, etc.) in the drop-down menu and the corresponding fragments and the structures will be displayed in the MPDS platform.

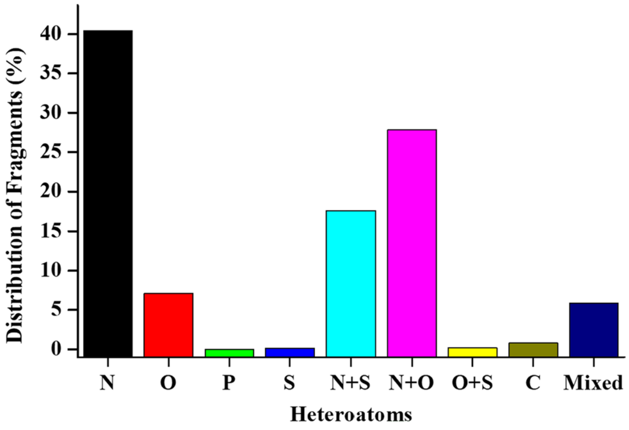


Fig. 5 Distribution of unique fragments based on heteroatoms

Conclusions

This work resulted in the development of ring fragments library with high diversity from the hetero-aliphatic ring structures. This library can allow the sampling of large chemical spaces comprising molecules with known biological activity, natural products, and theoretically generated molecules. The information provided from the analysis of the generated frsagment library can help in finding novel drug molecules which is a critical step for biomedical research. In this context, this library module is available in the MPDS web portal (http://mpds.neist.res.in:8085) to assist in molecule design, especially for computer-aided drug design. This module provides an inbuilt fragment library database to extract molecular fragments (classified as rings, linkers, and substituents), an open-source fragmentation tool, and a virtual screening tool. The fragment library in MPDS will be continuously updated in regular intervals with fragments from new chemical entities offering varied skeletons and scaffolds for designing molecules. We hope that this module will allow the scientific community to use the FBDD approach more effectively in the computational design of molecules.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11030-022-10506-5.

Acknowledgements GNS thanks DST for the award of the J C Bose National fellowship. Director, CSIR IICT is thanked for his support. DBT is thanked for the financial support in the form of Centre of Excellence in Advanced Computation and Data Sciences (Ref. No: BT/PR40188/BTIS/137/27/2021). CSIR is thanked for supporting the research in the form of project HCP41.

Declarations

Conflict of interest The authors have no conflict of interest.

References

1. Reddy AS, Priyadarshini S, Kumar PP, Pradeep HN, Sastry GN (2007) Virtual screening in drug discovery - a computational perspective. Curr Protein Pept Sci 8:329–351. https://doi.org/10.2174/138920307781369427
2. Bohari MH, Sastry GN (2012) FDA approved drugs complexed to their targets: evaluating pose prediction accuracy of docking protocols. J Mol Model 9:4263–4274. https://doi.org/10.1007/s00894-012-1416-1
3. Gaur AS, Bhardwaj A, Sharma A, John L, Vivek MR, Tripathi N, Bharatam PV, Kumar R, Janardhan S, Mori A, Banerji A, Lynn AM, Hemrom AJ, Passi A, Singh A, Kumar A, Muvva C, Madhuri C, Choudhury C, Kumar AD, Pandit D, Bharti DR, Kumar D, Singam AE, Raghava GPS, Sailaja H, Jangra H, Raithatha K, Tanneeru K, Chaudhary K, Karthikeyan M, Prasanthi M, Kumar N, Yedukondalu N, Rajput NK, Saranya PS, Narang P, Dutta P, Krishnan RV, Sharma R, Srinithi R, Mishra R, Hemasri S, Singh S, Venkatesan S, Kumar S, Jaleel UCA, Khedkar V, Joshi Y, Sastry GN (2017) Assessing therapeutic potential of molecules: molecular property diagnostic suite for tuberculosis (MPDS^{TB}). J Chem Sci 129:515. https://doi.org/10.1007/s12039-017-1268-4
4. Nagamani S, Gaur AS, Tanneeru K, Muneeswaran G, Madugula SS, MPDS Consortium, Druzhilovskiy D, Poroikov VV, Sastry GN (2017) Molecular property diagnostic suite (MPDS): development of disease-specific open-source web portals for drug discovery. SAR QSAR Environ Res 11:913–926. https://doi.org/10.1080/1062936X.2017.1402819
5. Gaur AS, Nagamani S, Tanneeru K, Druzhilovskiy D, Rudik A, Poroikov V, Sastry GN (2018) Molecular property diagnostic suite for diabetes mellitus (MPDS^{DM}): an integrated web portal for drug discovery and drug repurposing. J Biomed Inf 85:114–125. https://doi.org/10.1016/j.jbi.2018.08.003
6. Baker M (2013) Fragment-based lead discovery grows up. Nat Rev Drug Discov 12:5–7. https://doi.org/10.1038/nrd3926
7. Hoffer L, Renaud JP, Horvath D (2011) Fragment-based drug design: Computational and experimental state of the art. Comb Chem High Throughput Screen 14:500–520. https://doi.org/10.2174/138620711795767884
8. Ray PC, Kiczun M, Huggett M, Lim A, Prati F, Gilbert IH, Wyatt PG (2017) Fragment library design, synthesis and expansion: nurturing a synthesis and training platform. Drug Discov Today 22:43–56. https://doi.org/10.1016/j.drudis.2016.10.005
9. Badrinarayan P, Sastry GN (2012) Virtual screening filters for the design of type II P38 MAP kinase inhibitors: a fragment-based library generation approach. J Mol Graph Model 34:89–100. https://doi.org/10.1016/j.jmgm.2011.12.009
10. Bemis GW, Murcko MA (1996) The properties of known drugs. 1. Molecular frameworks. J Med Chem 39:2887–2893. https://doi.org/10.1021/jm9602928
11. Bemis GW, Murcko MA (1999) Properties of known drugs. 2. Side chains. J Med Chem 42:5095–5099. https://doi.org/10.1021/jm9903996
12. Congreve M, Carr R, Murray C, Jhoti HA (2003) 'Rule of Three' for fragment-based lead discovery? Drug Discov Today 8:876–877. https://doi.org/10.1016/s1359-6446(03)02831-9
13. Lewell XQ, Judd DB, Watson SP, Hann MM (1998) RECAP-retrosynthetic combinatorial analysis procedure: a powerful new technique for identifying privileged molecular fragments with useful applications in combinatorial chemistry. J Chem Inf Comput Sci 38:511–522. https://doi.org/10.1021/ci970429i
14. Degen J, Wegscheid-Gerlach C, Zaliani A, Rarey M (2008) On the art of compiling and using "drug-like" chemical fragment spaces. ChemMedChem 3:1503–1507. https://doi.org/10.1002/cmdc.200800178
15. Greenwood PE, Nikulin MS (1996) Wiley, New York. ISBN 0-471-55779-X
16. Morrison CN, Prosser KE, Stokes RW, Cordes A, Metzler-Nolte N, Cohen SM (2020) Expanding medicinal chemistry into 3D space: metallofragments as 3D scaffolds for fragment-based drug discovery. Chem Sci 11:1216–1225. https://doi.org/10.1039/c9sc05586j
17. Heikamp K, Zuccotto F, Kiczun M, Ray P, Gilbert IH (2018) Exhaustive sampling of the fragment space associated to a molecule leading to the generation of conserved fragments. Chem Biol Drug Des 91:655–667. https://doi.org/10.1111/cbdd.13129
18. Ghersi D, Singh M (2014) MolBLOCKS: decomposing small molecule sets and uncovering enriched fragments. Bioinformatics 30:2081–2083. https://doi.org/10.1093/bioinformatics/btu173
19. Liu T, Naderi M, Alvin C, Mukhopadhyay S, Brylinski M (2017) Break down in order to build up: decomposing small molecules for fragment-based drug design with *e*MolFrag. J Chem Inf Model 57:627–631. https://doi.org/10.1021/acs.jcim.6b00596
20. Li Y, Zhao Z, Liu Z, Su M, Wang R (2016) AutoT&T vol 2: an efficient and versatile tool for lead structure generation and optimization. J Chem Inf Model 56:435–453. https://doi.org/10.1021/acs.jcim.5b00691
21. Pevzner Y, Frugier E, Schalk V, Caflisch A, Woodcock HL (2014) Fragment-based docking: development of the CHARMMing web user interface as a platform for computer-aided drug design. J Chem Inf Model 54:2612–2620. https://doi.org/10.1021/ci500322k
22. Naderi M, Alvin C, Ding Y, Mukhopadhyay S, Brylinski M (2016) A graph-based approach to construct target-focused libraries for virtual screening. J Chem inform 8:1–6. https://doi.org/10.1186/s13321-016-0126-6
23. Fechner U, Schneider G (2007) Flux (2): comparison of molecular mutation and crossover operators for ligand-based de novo design. J Chem Inf Model 47:656–667. https://doi.org/10.1021/ci6005307
24. Kozakov D, Grove LE, Hall DR, Bohnuud T, Mottarella SE, Luo L, Xia B, Beglov D, Vajda S (2015) The FTMap family of web servers for determining and characterizing ligand-binding hot spots of proteins. Nat Protoc 10:733–755. https://doi.org/10.1038/nprot.2015.043
25. Tsai TY, Chang KW, Chen CYC (2011) IScreen: world's first cloud-computing web server for virtual screening and de novo drug design based on TCM Database@Taiwan. J Comput Aided Mol Des 25:525–531. https://doi.org/10.1007/s10822-011-9438-9
26. John L, Soujanya Y, Mahanta HJ, Sastry GN (2021) Chemoinformatics and machine learning approaches for identifying antiviral compounds. Mol Inform 23:2100190. https://doi.org/10.1002/minf.202100190
27. Madugula SS, John L, Nagamani S, Gaur AS, Poroikov VV, Sastry GN (2021) Molecular descriptor analysis of approved drugs using

unsupervised learning for drug repurposing. Comput Biol Med. https://doi.org/10.1016/j.compbiomed.2021.104856

28. Kumar N, Sharma H, Sastry GN (2021) Repurposing of approved drugs to predict new inhibitors for viral infectious diseases: a molecular modelling approaches. J Biomol Struct Dyn. https://doi.org/10.1080/07391102.2021.1905558
29. Kumar N, Sastry GN (2021) Study of lipid heterogeneity on bilayer membranes using molecular dynamics simulations. J Mol Graph Model. https://doi.org/10.1016/j.jmgm.2021.108000
30. Badrinarayan P, Sastry GN (2014) Specificity rendering 'hot-spots' for aurora kinase inhibitor design: the role of non-covalent interactions and conformational transitions. PLoS ONE. https://doi.org/10.1371/journal.pone.0113773
31. Badrinarayan P, Sastry GN (2011) Sequence, structure, and active site analyses of p38 MAP kinase: exploiting DFG-out conformation as a strategy to design new type II leads. J Chem Inf Model 51:115–129. https://doi.org/10.1021/ci100340w
32. Badrinarayan P, Sastry GN (2011) Virtual high throughput screening in new lead identification. Comb Chem High Throughput Screen 14:840–860. https://doi.org/10.2174/138620711797537102
33. Choudhury C, Priyakumar UD, Sastry GN (2016) Structural and functional diversities of the hexadecahydro-1H-cyclopenta[a]phenanthrene framework, a ubiquitous scaffold in steroidal hormones. Mol Inform 35:145–157. https://doi.org/10.1002/minf.201600005
34. Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H (2007) The scaffold tree visualization of the scaffold universe by hierarchical scaffold classification. J Chem Inf Model 47:47–58. https://doi.org/10.1021/ci600338x
35. Hoffmann T, Gastreich M (2019) The next level in chemical space navigation: going far beyond enumerable compound libraries. Drug Discov Today 24:1148–1156. https://doi.org/10.1016/j.drudis.2019.02.013
36. Gaulton A, Hersey A, Nowotka M, Bento AP, Chambers J, Mendez D, Mutowo P, Atkinson F, Bellis LJ, Cibrián-Uhalte E, Davies M (2017) The ChEMBL database in 2017. Nucleic Acids Res 45:945–954. https://doi.org/10.1093/nar/gkw1074
37. Ursu O, Holmes J, Bologa CG, Yang JJ, Mathias SL, Stathias V, Nguyen DT, Schürer S, Oprea T (2019) DrugCentral 2018: an update. Nucleic Acids Res 47:963–970. https://doi.org/10.1093/nar/gky963
38. Liu Z, Li Y, Han L, Li J, Liu J, Zhao Z, Nie W, Liu Y, Wang R (2015) PDB-wide collection of binding data: current status of the PDBbind database. Bioinformatics 31:405–412. https://doi.org/10.1093/bioinformatics/btu626
39. Sorokina M, Merseburger P, Rajan K, Yirik MA, Steinbeck C (2021) COCONUT online: collection of open natural products database. J Cheminform 13:1–3. https://doi.org/10.1186/s13321-020-00478-9
40. Patel H, Ihlenfeldt WD, Judson PN, Moroz YS, Pevzner Y, Peach ML, Delannée V, Tarasova NI, Nicklaus MC (2020) SAVI, in silico generation of billions of easily synthesizable compounds through expert-system type rules. Sci Data 7:1–4. https://doi.org/10.1038/s41597-020-00727-4
41. Douguet D (2010) E-LEA3D: A computational-aided drug design web server. Nucleic Acids Res 38:615–621. https://doi.org/10.1093/nar/gkq322
42. Hao GF, Jiang W, Ye YN, Wu FX, Zhu XL, Guo FB, Yang GF (2016) ACFIS: a web server for fragment-based drug discovery. Nucleic Acids Res 44:550–556. https://doi.org/10.1093/nar/gkw393
43. Landrum G (2016) Rdkit: open-source cheminformatics software, https://github.com/rdkit/rdkit, 149:150.
44. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR (2011) Open Babel: an open chemical toolbox. J Cheminform 3:1–4. https://doi.org/10.1186/1758-2946-3-33
45. Sander T, Freyss J, von Korff M, Rufener C (2015) DataWarrior: an open-source program for chemistry aware data visualization and analysis. J Chem Inf Model 55:460–473. https://doi.org/10.1021/ci500588j
46. Ritchie TJ, Macdonald SJ (2009) The impact of aromatic ring count on compound developability—are too many aromatic rings a liability in drug design? Drug Discov Today 14:1011–1020. https://doi.org/10.1016/j.drudis.2009.07.014
47. Hou TJ, Xia K, Zhang W, Xu XJ (2004) ADME evaluation in drug discovery. 4. Prediction of aqueous solubility based on atom contribution approach. J Chem Inf Comput Sci 44:266–275. https://doi.org/10.1021/ci034184n

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.