Molecular Diversity (2024) 28:3243–3259
https://doi.org/10.1007/s11030-023-10752-1

ORIGINAL ARTICLE

Molecular Property Diagnostic Suite Compound Library (MPDS-CL): a structure-based classification of the chemical space

Lijo John[1,2] · Selvaraman Nagamani[1,2] · Hridoy Jyoti Mahanta[1,2] · S. Vaikundamani[1] · Nandan Kumar[1,2] · Asheesh Kumar[1] · Esther Jamir[1,2] · Lipsa Priyadarsinee[1,2] · G. Narahari Sastry[1,2]

Received: 5 August 2023 / Accepted: 17 October 2023 / Published online: 30 October 2023

Abstract

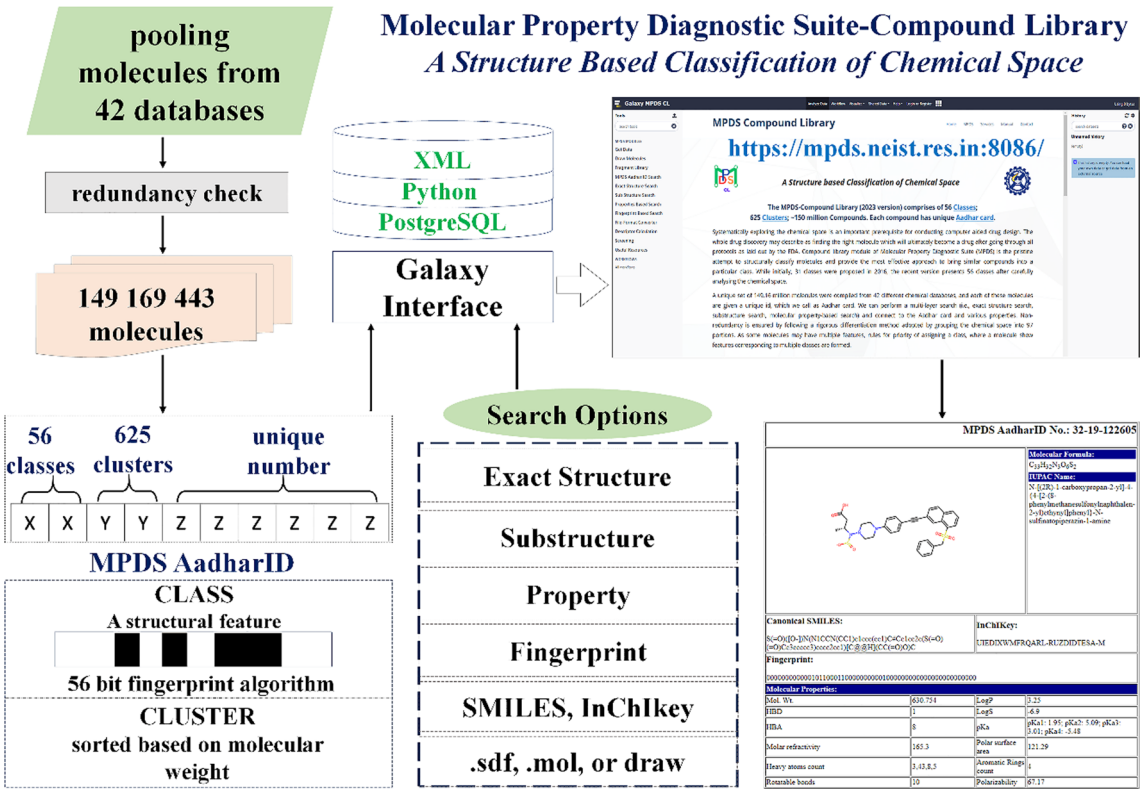
Molecular Property Diagnostic Suite Compound Library (MPDS-CL) is an open-source Galaxy-based cheminformatics web portal which presents a structure-based classification of the molecules. A structure-based classification of nearly 150 million unique compounds, obtained from 42 publicly available databases and curated for redundancy removal through 97 hierarchically well-defined atom composition-based portions, has been done. These are further subjected to 56-bit fingerprint-based classification algorithm which led to the formation of 56 structurally well-defined classes. The classes thus obtained were further divided into clusters based on their molecular weight. Thus, the entire set of molecules was put into 56 different classes and 625 clusters. This led to the assignment of a unique ID, named as MPDS-AadharID, for each of these 149,169,443 molecules. MPDS-AadharID is akin to the unique number given to citizens in India (similar to SSN in the US and NINO in the UK). The unique features of MPDS-CL are (a) several search options, such as exact structure search, substructure search, property-based search, fingerprint-based search, using SMILES, InChIKey and key-in; (b) automatic generation of information for the processing for MPDS and other galaxy tools; (c) providing the class and cluster of a molecule which makes it easier and fast to search for similar molecules and (d) information related to the presence of the molecules in multiple databases. The MPDS-CL can be accessed at https://mpds.neist.res.in:8086/.

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

Graphical abstract



Keywords Chemical space · Structural classification · Compound library · MPDS-AadharID

Abbreviations

AI	Artificial Intelligence
AWK	Aho, Weinberger, and Kernighan
BCS	Biopharmaceutical Classification System
CDK	Chemistry Development Kit
Da	Dalton
FAIR	Findability, Accessibility, Interoperability and Reusability
FTP	File Transfer Protocol
GPCR	G Protein Coupled Receptors
GRAC	Guide to Receptors and Channels database
HMDB	Human Metabolome Database
InChI	International Chemical Identifier
KEGG	Kyoto Encyclopaedia of Genes and Genomes
MOL	Molecular format
MPDS-CL	Molecular Property Diagnostic Suite Compound Library
$MPDS^{COVID-19}$	Molecular Property Diagnostic Suite-Coronavirus Disease-19
$MPDS^{HIV}$	Molecular Property Diagnostic Suite-Human Immunodeficiency Virus
$MPDS^{NAFLD}$	Molecular Property Diagnostic Suite-Non-Alcoholic Fatty Liver Disease
$MPDS^{TB}$	Molecular Property Diagnostic Suite-Tuberculosis
$MPDS^{DM}$	Molecular Property Diagnostic Suite-Diabetes Mellitus
MW	Molecular weight
NCI	National Cancer Institute
NINO	National Insurance Number
PharmGKB	Pharmacogenomics Knowledgebase
QSAR	Quantitative Structure–Activity Relationship
SCONP	Structural Classification of Natural Products
SDF	Structured Data File
SED	Stream Editor
SHA-256	Secure Hash Algorithm 256-bit
SMARTS	SMILES arbitrary target specification
SMILES	Simplified Molecular Input Line Entry System
SMPDB	Small Molecule Pathway Database
SSN	Social Security Number
UK	United Kingdom

US	United States
XML	Extensible Markup Language

Introduction

Chemical space is quite vast and finding a molecule with the desired property is arguably the most formidable challenge. In general, structurally similar compounds are expected to have similar properties. In drug/molecular design, the structural similarity is of paramount importance and any effort which structurally and systematically divides the chemical space will be of outstanding interest [1–5]. Developing new chemical libraries is of fundamental importance in the current scenario to systematize the process of covering huge chemical space and tapping its potential in multifarious applications in science and technology [6–10]. Such an effort will help to qualitatively and quantitatively estimate and assess the structural similarity and chemical diversity which will be useful in mining the chemical/biological property space [11–13]. The ability to synthesize molecules has remarkably enhanced due to the pioneering efforts by experimental chemists, which resulted in the synthesis of a huge number of molecules of diverse structural scaffolds and features [1]. However, in practice only a very small fraction of such synthesized molecules is of utility, which highlights the limitations of exploratory approaches and emphasizes the need to adopt rational design. Therefore, in recent years, the focus has shifted from "how to synthesize" to "what to synthesize".

While there are extensive studies on chemical space, most of them are devoted to explore the property space and very few of them focus on structure-based classification. One of the pioneering attempts was made by Waldmann's group in attempting to a Structural Classification of Natural Products (SCONP), which is thus limited to only natural products [4, 14]. The focus was on heterocycles and the occurrence of those compounds in natural products. Other approaches are based on the theoretical generation of molecules, and explore the size of the chemical space [1–5, 15–18]. Fragment-based approaches also have played a significant role, and several methods were developed especially in the area of medicinal chemistry-directed drug discovery [19–21].

Compound libraries are developed with the objective of enumeration, analysis, and extrapolation of the chemical space for various applications in chemistry, biology, and allied sciences. The curation of the chemical data are also concerned with the cleaning of molecules to remove any salts, and mixtures, normalization of various chemotypes, de-duplication of redundant molecules, etc. Besides, the manual and automated curation applied to the big chemical data, the lack of rigorous standardization methods in the chemical reaction, transformations are still one of the problems faced by chemists and the role of informatics and Artificial Intelligence (AI) is valuable in removing the barriers and deriving novel insights from the vast molecular space [22–28].

Finding druglike and non-druglike molecules through various means of theory and experimentation is the prime motto of drug discovery projects. While there are a large number of databases depicting the chemical structures, to our knowledge, attempts towards the structural classification of compounds are scarce. The recent progress made in the synthesis and the growing need for novel chemical entities together pushes for an urgent need to scale up the existing methods and design new methods in developing elegant technologies for making the best use of deciphering the structure–property relationships from the chemical space [8–12]. The ultimate goal in all these attempts is to find the molecule(s) with the most desirable properties, e.g., drugs, catalysts, agrochemicals, etc. [13, 14].

The Galaxy-based MPDS was an initiative to strengthen the open-source computational drug discovery, providing access to most of the available open-source, custom designed indigenously developed scripts, programs and software packages [29–34]. Galaxy platform supports both the web and the standalone version which can be implemented on a Linux server. The Toolshed of Galaxy, which is periodically updated, is populated with a wide range of programs that can be directly imported and installed on users' Galaxy portals [30, 31, 35]. It also offers the advantage of adding several user-developed programs which are incorporated into the Galaxy directory and are programmatically called to the front interface through an XML file. Several virtual machine images of Galaxy instances [30] are also made available online so that they can be used for various hands-on and other training sessions for data-intensive biology and chemistry applications. MPDS-CL in specific and other chemical libraries in general; its development, scalability, and automation techniques will be essential in deriving novel insights for drug discovery by comprehensively assessing the chemical space and finding various ways of prioritizing the lead molecules for drug discovery projects [36–40]. Ensuring the unique molecules along with creating a structurally well-defined chemical data library was taken as the paramount significance in creating the MPDS-CL.

Genesis of Galaxy-based MPDS

MPDS is an indigenous initiative that is developed to strengthen computational drug discovery and is an attempt to address the pressing issues of drug discovery. As it is developed on the Galaxy platform, features like cloud-based accessibility, reproducibility, and various data-driven methods are also made available. The MPDS suites of disease-specific web portals include proprietary libraries, machine

learning models, and other relevant open-source tools and resources essential for drug discovery solutions. The compound library is the most important component of the data library module of MPDS, which was initially integrated into the MPDS^{TB} web portal. At that time, using the six most popular and abundant chemical databases, around 110 million unique non-redundant set of compounds, with 31 classes, was reported. Over a period of time, various disease-specific MPDS portals including MPDS^{TB} [15], MPDS^{DM} [34], MPDS$^{COVID\text{-}19}$ [41] were developed, and other portals like MPDS^{NAFLD} and MPDS^{HIV} are under development. The modules in MPDS are categorized into (i) data library (that consists of information on genes and targets specific to a disease or disorder, a molecular repository, druglike fragments, literature, etc.,) (ii) data processing (computation of molecular descriptors/fingerprints, file format converter) (iii) data analysis (QSAR, docking, drug-likeness filter, and visualization tool) and (iv) Advanced modules (which include various predictive modules for disease–disease interaction, big data analysis, and machine learning tools). MPDS is also equipped with a workflow management system that enables the users to easily integrate multiple tools from the available modules and customize the existing workflows as per the requirements [32–34, 41]. The utility and scope of open-source packages are well-documented in the literature [42, 43].

The current MPDS-CL was developed into a new full-fledged web portal, and not as one of the modules of MPDS and in that respect it is vastly different from the earlier module, which was presented as one of the modules of MPDS^{TB} about 6 years ago. The MPDS-CL is an independent web portal, which is well-positioned to integrate with Galaxy and MPDS web portals. The classification of close to 150 million molecules and redundancy removal techniques employed are different and much more efficient, compared to those employed in MPDS^{TB} six years ago (Fig. 1). This module enables comprehensive structural analysis, assigns a unique MPDS-AadharID to each molecule, offers descriptor analysis tools, fragment library, and screening tools.

Materials and methods

Forty two public domain chemical databases, have been considered in making the current compound library (Table 1), while the erstwhile compound library module of MPDS^{TB} had six databases (PubChem, KEGG, ZINC, DrugBank, ASINEX, and NCI). However, two types of databases were excluded: (a) large databases of hypothetical molecules and (b) some commercial/inaccessible databases (Table S1).

The databases considered here may be categorized as general and specialized databases based on the type of molecules they contain. The bioactivity libraries constitute repositories like PubChem and ChEMBL. While other categories of databases include drugs and molecules of biological importance such as Therapeutic Target Database, DrugCentral, SuperDrug2, DrugBank, PharmGKB, GRAC, SMPDB, KEGG compound database, HMDB, and ChEMBL-DNDi. Other libraries include molecules extracted from patent and general literature (SureChEMBL, BindingDB), GPCR ligand database, GPCR decoy database, a database of lipid-like molecules (LipidBank), a database of

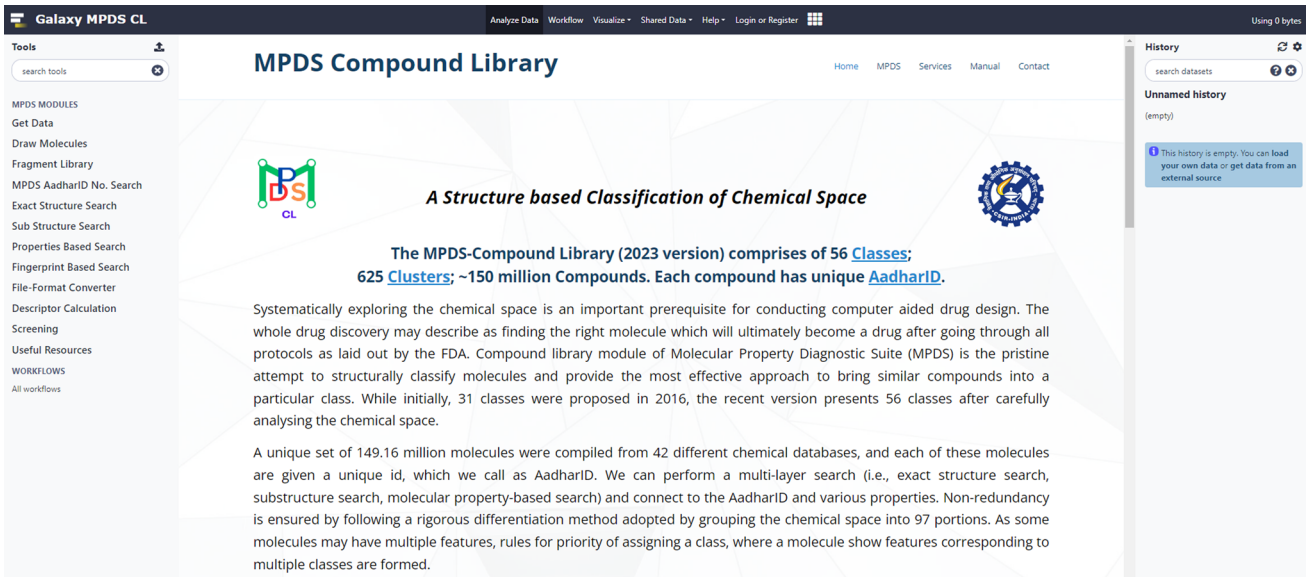


Fig. 1 Home page of MPDS Compound Library (MPDS-CL) accessible at https://mpds.neist.res.in:8086/. The left panel consists of various search options and cheminformatics tools incorporated in the portal. The right-side panel displays the uploaded data and the results

Table 1 Detailed information of the databases used to develop the MPDS-CL with their statistics

References	Database	Dec-16 Compounds	Dec-17	May-18	May-19	Jun-20	Jul-22	Dec-22	Jul-23
[61]	PubChem	92,293,546	94,136,411	96,300,363	97,176,562	102,710,207	111,665,090	112,428,952	114,022,856
[62]	Mcule	–	–	35,742,734	42,267,878	45,472,755	85,383,043	85,383,043	
[63]	eMolecules (free version)	–	–	17,559,951	22,327,838	26,436,139	26,436,139	26,436,139	26,436,139
[64]	SureChEMBL	16,599,522	18,971,423	20,255,239	20,926,618	21,574,903	21,641,384	> 21,641,384	> 21,641,384
[65]	CoCoCo	6,981,556	6,981,556	6,981,556	6,981,556	6,981,556	6,981,556	6,981,556	6,981,556
[66]	ChEMBL	1,686,695	1,686,695	1,828,820	1,879,206	1,950,765	2,331,200	2,331,200	3,379,776
[67]	ChemDiv	–	–	1,600,000	1,600,000	1,600,000	1,600,000	1,600,000	1,600,000
[68]	SPECS	–	–	1,024,181	1,361,884	1,361,884	1,361,884	1,361,884	1,361,884
[69]	ChEMBL-DNDi	1,305,058	1,305,058	1,305,058	1,305,058	1,305,058	1,305,058	1,305,058	1,305,058
[70]	Ligand-Info	206,334	206,334	1,159,274	1,159,274	1,159,274	1,159,274	1,159,274	1,159,274
[71]	GPCR Decoy DB	980,655	980,655	980,655	980,655	980,655	980,655	980,655	980,655
[72]	BindingDB	–	–	650,012	652,068	820,433	1,098,225	1,098,225	1,598,620
[73]	TimTec	–	–	628,462	628,462	628,462	628,462	628,462	628,462
[74]	ASINEX	596,300	603,276	610,548	610,548	610,548	610,548	610,548	610,548
[75]	InterBioScreen	–	–	564,386	569,704	569,704	569,704	569,704	569,704
[76]	COCONUT	–	–	–	–	–	407,270	407,270	407,270
[77]	Universal Natural Products Database	298, 716	298,716	298,716	298,716	298,716	298,716	298,716	298,716
[78]	NCI	284,176	284,176	284,176	284,176	284,176	284,176	284,176	284,176
[79]	Crystallography Open DB	–	–	162,172	162,172	189,067	491,597	491,597	491,597
[80]	HMDB	41,824	114,100	114,100	114,100	114,100	251,936	251,936	251,936
[81]	Probes and Drug portal	–	–	–	–	–	–	98,657	158,238
[82]	Openmolecules	–	–	90,155	90,155	90,155	90,155	90,155	90,155
[83]	LipidBank	–	–	84,112	84,112	84,112	84,112	84,112	84,112
[84]	ChemBank	–	–	75,964	75,964	75,964	75,964	75,964	75,964
[85]	ChemMine	–	–	64,000	64,000	64,000	64,000	64,000	64,000
[86]	ChEBI	50,504	53,495	54,724	55,453	55,453	60,176	60,176	103,694
[87]	SMP database	–	–	4341	49,817	49,817	49,817	49,817	99,607
[88]	GPCR Ligand	25,145	25,145	25,145	25,145	25,145	25,145	25,145	25,145
[89]	KEGG	7913	18,111	18,228	18,228	18,228	18,228	18,228	18,228
[90]	MDPI	8576	8576	15,348	15,348	15,348	22,401	22,401	22,401
[46]	OSADHI	–	–	–	–	–	–	22,314	22,314
[91]	TOSLab	17,410	17,410	17,410	17,410	17,410	17,410	17,410	17,410
[92]	DrugBank	7002	10,507	11,146	11,924	11,924	14,665	14,665	15,483
[93]	MyriaScreen2	10,000	10,000	10,000	10,000	10,000	10,000	10,000	10,000
[94]	GRAC	–	–	6881	9405	9405	9405	9405	9405
[45]	NEI-MPDB	–	–	–	–	–	–	9225	9225
[95]	PHARMGKB	–	–	7030	7030	7030	7030	7030	7030
[96]	ChemDB	–	–	5937	5937	5937	5937	5937	5937
[97]	DrugCentral	–	–	–	3981	4052	4099	4099	8376
[98]	SuperDrug2	–	–	3982	3993	3993	3993	3993	3993
[99]	TTD	–	–	2326	2936	2936	2936	2936	26,316
[100]	ZINC[a]	–	–	1 billion	1 billion	> 2 billion	> 2 billion	> 2 billion	> 2 billion
[2]	SAVI[b]	–	–	–	–	1.75 billion	1.75 billion	1.75 billion	1.75 billion

[a]A specific subsets (only the "Standard + In-Stock" subset) has been included in MPDS-CL

[b]Not included in the MPDS-CL

the crystal structure of organic and inorganic, metal-organics and other minerals (Crystallography-open database), Natural product database (InterBioScreen, UNPD), and database with lead-like molecule and useful fragments (ASINEX and ChemDiv) (Table 1). The methodology described in the further sections essentially deals with various open-source cheminformatics packages and the Python programs developed for chemical data analysis.

Retrieval of molecules

The molecules used for developing the compound library were retrieved from various public domain chemical databases in multiple file formats like SMILES, SDF, and MOL. The molecules that were not in the SMILES format, were converted to the canonical SMILES to maintain a unique representation for all the molecules in the database. The canonicalization algorithm as implemented in the OpenBabel3 [44] was used to convert all the SMILES formats into the canonical format. Few of these databases are updated over a specific period while others like PubChem, SureChEMBL are populated with new molecules every other day. In the current study, we have retrieved the molecules till July 2023 and all analysis was performed using this dataset. The chemical databases offer a wide range of information about the molecules including their structural, physicochemical, reaction profiles, analytical data, etc. The molecules that were obtained from public domain chemical databases are mentioned in Table 1 in which each database and its statistics of update is indicated. The work summarized here mainly consists of the structural information in the form of SMILES, which were parsed into the SMARTS pattern for subsequent processing and analysis of the dataset. Linux-based expressions, such as AWK, sed, along with a set of python packages were employed to obtain unique, non-redundant 97 atom-based portions (Table S2). The process of retrieving data from each database differs and retrieval depends on the type and number of new molecules included in the database. In the case of PubChem, molecules included within a specific duration were retrieved using the file transfer protocol (FTP) method whereas for ChEMBL, a web resource client that is a python-based library, as well as a bulk download option, was used. In the case of SureChEMBL, the quarterly updated molecules were retrieved in bulk. ZINC database has the option to retrieve specific subsets, "Tranch" according to the molecule's type, its reactivity, purchasability, etc., While there are a large number of molecules in ZINC, we considered only the "Standard + In-Stock" subset. From all the remaining databases, the molecules were retrieved as 'bulk download (Fig. 2). Some region-specific natural products and phytochemicals-based databases are also available and some of them are developed in our group [45, 46], and if new molecules are found they will be added to MPDS-CL periodically, during the half-yearly updates.

Schema for redundancy removal and structural classification

Each database was classified into 97 hierarchically well-defined atom-based portions (Fig. 3). As the classification is hierarchical, those portions which were not considered for further classification were labelled as 'terminal portions', and the rest as 'open portions'. The classification has been carried out in five distinct classification steps, which are called as layers (Fig. 3). The first layer of classification was based on molecular weight (MW) and atom composition. Thus, Portions 1–3 were assigned as terminal portions for the first classification layer, and all remaining portions were further classified in the second layer. The second classification layer was based on molecular topology (acyclic or cyclic), and all those molecules that were categorized as acyclic were considered as terminal portions, while cyclic molecules were further classified as alicyclic and aromatic molecules in the third layer.

In the third classification layer, both alicyclic and aromatic portions that contain Te, Se, Ge, As, Sb (Portions 12–13), B and Si (Portions 14–19), Phosphorus (Portions 20–21), and hydrocarbons (Portions 22–25), were assigned as terminal portions, and the remaining portions were considered for further classification. The fourth classification layer is based on the count of heteroatoms, which generate portions that were specifically categorized as sulphur, oxygen, and nitrogen-containing. These were subjected to further classification as no terminal portions were identified in this layer. The open portions obtained from the fourth layer were used for classification in the fifth layer, and it was based on the position of the heteroatom in a molecule that can be inside or outside the ring. For example, if a molecule consists of sulphur inside a ring, then it was classified into a separate portion, and if the molecule doesn't have any sulphur atom inside the ring, it was classified into a different portion. Likewise, the oxygen and nitrogen-containing molecules were classified into their respective portions. In this way, all the open portions of the fourth layer were completely classified as terminal portions in the fifth layer, and hence a total of 97 different atom-based portions were obtained.

Among the file formats, InChIKey [47] appears to be the gold standard for unambiguously identifying the molecule, it has been used for redundancy removal in the compound library. The standard InChIKey was computed using OpenBabel3 [44] for all the molecules. InChIKey is a string of 27 characters built on the SHA-256 encoding algorithm applied on the InChI and the abstract notation consisting of hashed information for molecular skeleton and isomerism. It was

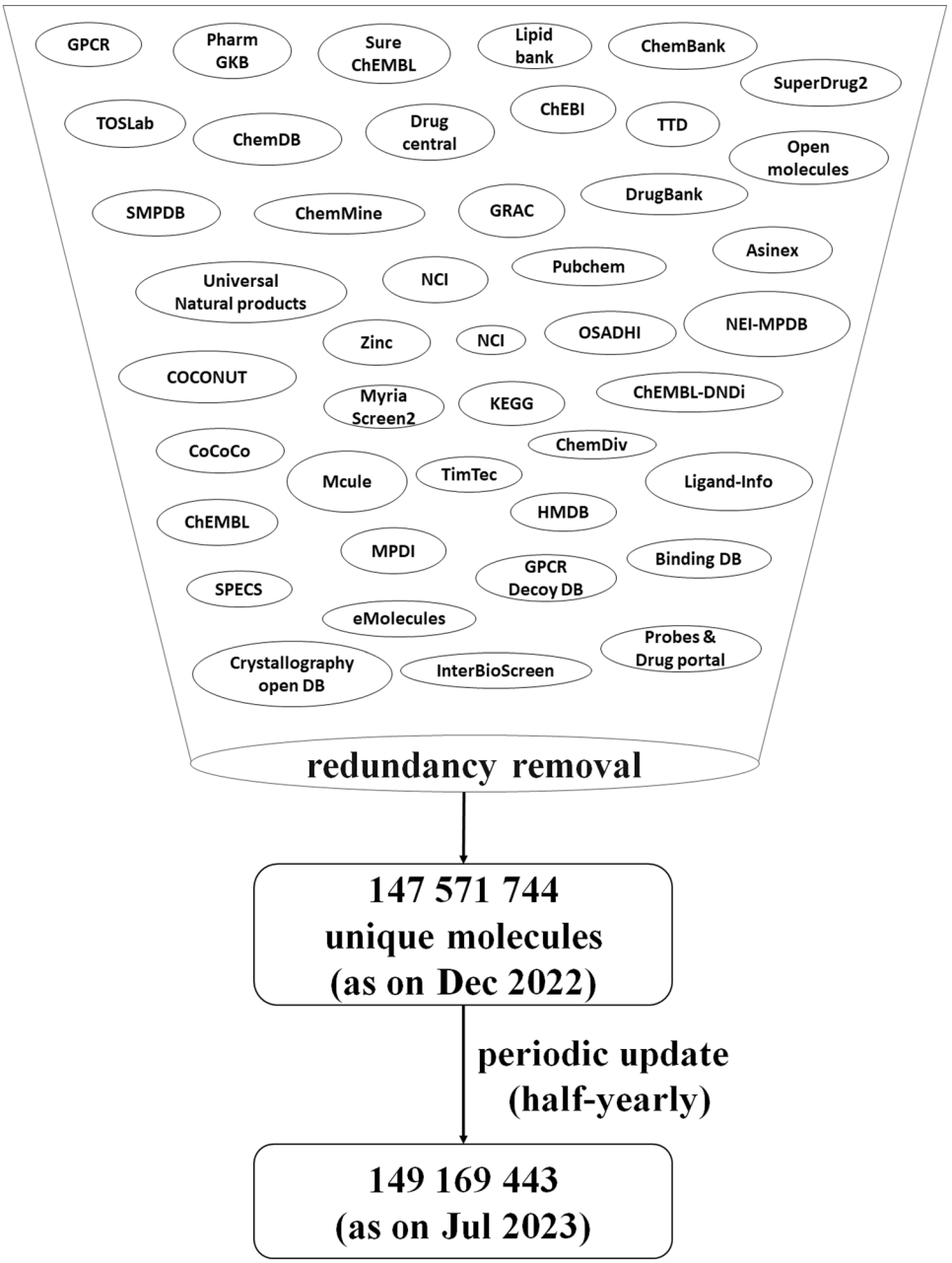


Fig. 2 An overview of the development of the MPDS-CL, from the molecules downloaded from 42 databases. MPDS-CL had 147,571,744 molecules in December 2022. While the initial set of non-redundant molecules were obtained by employing the scripts in portions, the half-yearly updates use an entirely different approach to add molecules directly to the classes and clusters. Thus, the updates as on July 2023 is 149,169,443 number of unique molecules

primarily developed for indexing purpose but is very useful for text-based mining and searching in chemical databases, as well as for removing redundancy.

All programs used to classify the molecular space were coded in Python3 and SMARTS patterns were utilized extensively [48]. The validation of SMARTS patterns was done rigorously to avoid any misclassification of molecules. For atoms-based classification, the SMARTS pattern consisting of the acronym of individual atoms or corresponding atomic number was used, while the next level of classifying the cyclic and acyclic molecules was done by identifying the atoms in rings and outside the rings. The structural classification of molecules was primarily carried out on all the cyclic portions to identify the structural diversity of molecules. The algorithm used to classify the molecules first parses the SMILES string and converts it to the SMARTS pattern by calling the 'pybel function'of the OpenBabel package. This pattern is then searched in the parsed SMILES, and based on the presence/absence or position of specific atoms, molecules are classified (Fig. 4). The programs developed for structural classification have extensively made use of the SMARTS pattern parsed through both OpenBabel and RDkit modules [49], and individual SMARTS-based expression was created and validated for each class. Each of these SMARTS patterns act as substructure queries for

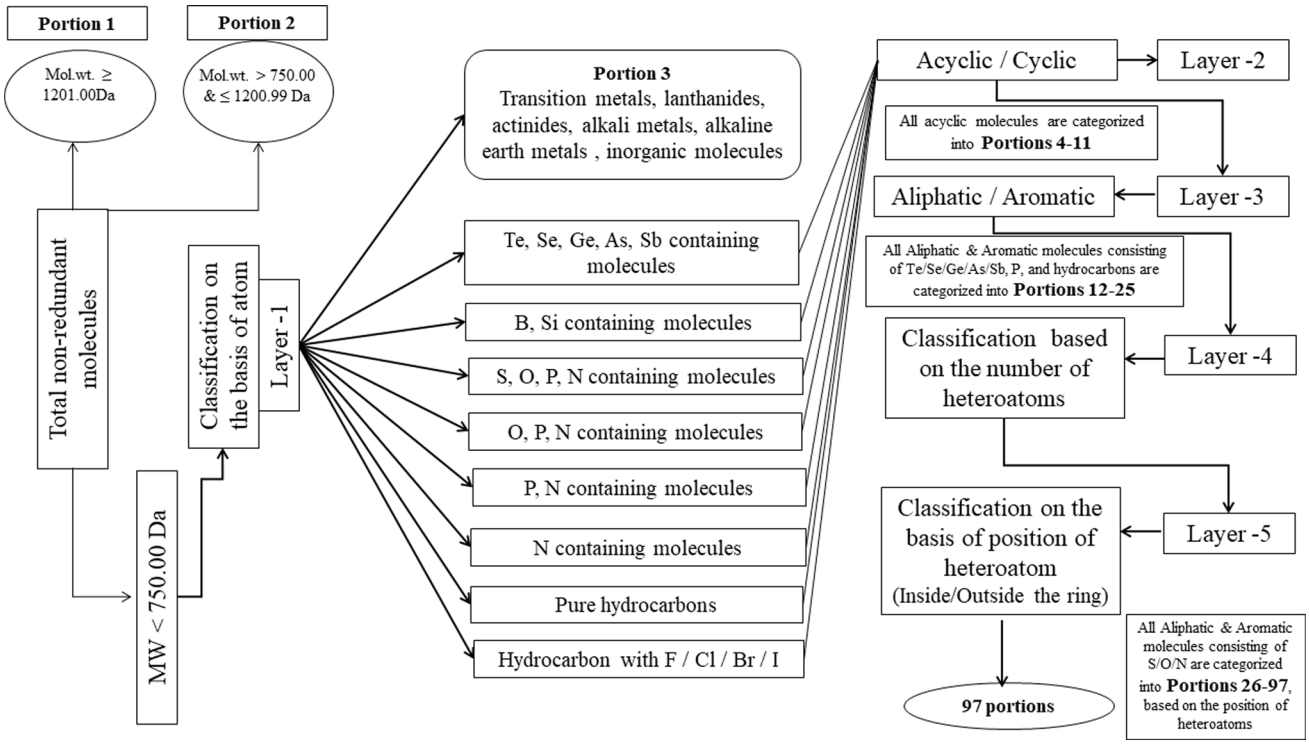


Fig. 3 Diagrammatic representation of the scheme employed to obtain the atom-based division of molecules into 97 portions

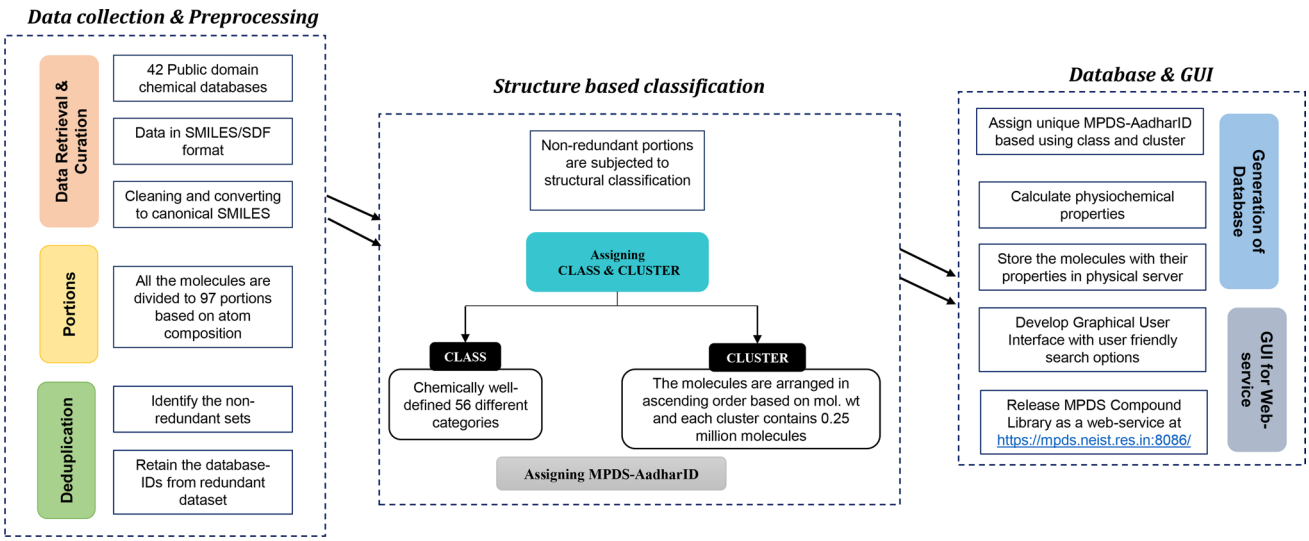


Fig. 4 An illustration depicting the step-by-step protocol employed for developing the MPDS-CL

identifying specific structural feature in a molecule from the large chemical space, and thus aiding in classifying the molecules with matched features. Particular syntax such as presence of rings, 'n-'ring-membered type, ring systems (fused/non-fused/connected) and other complex structural features, all were used to screen and efficiently identify molecules from large space. Considering the responsibility of producing and reusing scholarly data, we have strictly complied with the FAIR Data Principles which are well accepted, concise and measurable [50].

Results and discussion

After obtaining an unambiguous set of 149,169,443 molecules, structural classification of them into 56 classes was achieved by employing a series of scripts and tools, as described in the foregoing sections (Table 2). The 56 classes presented in the Table 2, are carefully carved to group and enumerate molecules with such features into distinguishable groups.

One needs to consider that a molecule may belong to multiple classes, but the sole idea of structural classification is to identify or designate a molecule on the basis of the core moiety or a principal structural feature responsible for rendering a specific activity, that again depends on its structure–property-based relationship. In literature various studies were reported [1–10] which focus on various hierarchical levels of structural classification such as topological, skeleton, atomic connectivity, formulae, and biological role. However, the current classification scheme is based on identifying and profiling molecules based on a chemically well-defined structural motif, termed as a '*class*'.

In this manuscript, the entire molecular structure was considered for enumerating various types of molecules existing in this portion of the chemical space. Two cyclic molecules combine in various ways, (a) edge sharing, which is called fused, (b) vertex sharing, which is spiro, (c) connected by a bond, which may be called connected, or (d) connected by a linker, which are called disconnected (Fig. 5). As seen in Table 2, Class 21 (Bicyclic connected rings) is found to have the largest population of molecules, i.e., 30.11 million, which clearly indicates that this class consists of the large number of molecules that have ring connected via a non-ring bond and absence of fused ring systems. The presence of connected rings systems offers stability and rigidity to the molecules in addition which may be responsible for their higher synthetic accessibility. Classes 10–12, 16–17, and 27 were designed to group specific molecules where free forms of Pyrrole, Furan, Thiophene, Benzene, and Pyridine can be identified. This is also done with an interest in identifying molecules with medicinally relevant rings exhibiting their unique functional role as valuable building block in drug design, optimizing certain therapeutic effects with respect to a drug, etc. Classes 22–37 are different forms of bicyclic fused ring systems, which are intended to group molecules with varied bicyclic scaffolds in terms of ring size, and combination of aliphatic and aromatic patterns. These classes combinedly constitute 25.85 million molecules, showing their immense contribution to the therapeutic chemical space, with a large population representing the terpenes, alkaloids, and other molecules belonging to

Table 2 List of 56 classes and population of molecules belonging to each class

Class	Description	Population
1	Acyclic (saturated/unsaturated)	3,455,531
2	Pure inorganic molecules	12,630
3	Monocyclic 3 membered saturated ring	333,466
4	Monocyclic 3 membered unsaturated ring	5711
5	Monocyclic 4 membered saturated ring	249,354
6	Monocyclic 4 membered unsaturated ring	9553
7	Monocyclic 5 membered saturated ring	1,224,146
8	Monocyclic 5 membered unsaturated ring	153,678
9	Monocyclic 5 membered aromatic ring	2708
10	Pyrrole (free)	171,154
11	Furan (free)	288,279
12	Thiophene (free)	604,404
13	Monocyclic 6 membered saturated ring	2,452,169
14	Monocyclic 6 membered unsaturated ring	405,431
15	Monocyclic 6 membered aromatic ring	77,290
16	Benzene (free)	26,496,311
17	Pyridine (free)	5,221,873
18	Monocyclic ≥ 7 membered saturated ring	620,637
19	Monocyclic ≥ 7 membered unsaturated ring	118,401
20	Multiple (≥ 2) main group elements in a ring	16,304,254
21	Bicyclic connected rings	30,111,865
22	Bicyclic fused 3 + 'n' membered	211,041
23	Bicyclic fused 4 + 'n' membered	365,491
24	Bicyclic fused 5 + 5 membered [A + A]	216,117
25	Bicyclic fused 5 + 5 membered [A + NA]	168,626
26	Bicyclic fused 5 + 6 membered [A + A]	5,300,636
27	Indole (Free)	1,631,787
28	Bicyclic fused 5 + 6 membered [A + NA]	1,079,827
29	Bicyclic fused 5 + 6 membered [NA + A]	3,620,404
30	Bicyclic fused 5 + (5/6/ ≥ 7) membered [NA + NA]	1,962,058
31	Bicyclic fused 5 + ≥ 7 membered [A + NA]	167,017
32	Bicyclic fused 6 + 6 membered [A + A]	4,729,630
33	Bicyclic fused 6 + 6 membered [A + NA]	3,658,454
34	Bicyclic fused 6 + 6 membered [NA + NA]	885,091
35	Bicyclic fused 6 + ≥ 7 membered	709,655
36	Bicyclic fused ≥ 7 + ≥ 7 membered	16,923
37	Bicyclic spiro	1,133,303
38	Tricyclic connected (1 ring aromatic)	715,276
39	Tricyclic connected (2 rings aromatic)	3,448,581
40	Tricyclic connected (3 rings aromatic)	1,115,644
41	Tricyclic connected (no aromatic rings)	52,924
42	Tricyclic fused (1 ring aromatic)	809,585
43	Tricyclic fused (2 rings aromatic)	1,257,912
44	Tricyclic fused (3 rings aromatic)	609,406
45	Tricyclic fused (no aromatic rings)	579,092
46	Tricyclic fused-connected [Any combinations]	11,066,401
47	Tetracyclic connected [Any combinations]	873,036
48	Tetracyclic fused [Any combinations]	1,167,417

Table 2 (continued)

Class	Description	Population
49	Tetracyclic fused + connected [Any combinations]	4,559,512
50	Complex ring systems up to tetracyclic	1,013,766
51	Pentacyclic & above	3,962,984
52	1 transition metal in a molecule	56,796
53	2 transition metals in a molecule	1871
54	≥ 3 transition metals in a molecule	303
55	Mol.wt. = 750.00–1200.99 Da	2,973,899
56	Mol.wt. ≥ 1201.00 Da	730,133
Total		149,169,443

A aromatic, *NA* non aromatic

pharmaceutically important natural products. The tricyclic molecules are grouped from classes 38–46, each class depicting the molecules with notable topological variations in association with different ring systems. Tetracyclic and polycyclic classes are arranged from 47 to 51, with a special class designated for large complexed fused rings up to tetracyclic group in class 50, while classes 52–54 are exclusively designed for molecules constituting transition metals, and classes 55–56 are designated for large MW (≥ 750.00 Da) molecules.

Table 3 illustrates the various possible combinations as explained between the two cyclic moieties, by taking a simple ring system. Molecules which contain both five and six membered rings simultaneously represent the largest chunk, 66.16 million, which represents more than 44% of all chemical space. Therefore, four out of 10 molecules have both 5 and 6 membered rings in them. The next best combination is three and six membered rings with a total of 2.59 million molecules. Among the classes, three or more transition metal containing molecules class 54, with a population of only 303 represent the least abundant class in the chemical space. The least population (1445) of molecules is obtained from those constituting both eight and ≥ 9 membered ring systems.

Property space for cyclic and acyclic molecules

While the molecules are classified as cyclic and acyclic systems, the general property distribution in the chemical space between these varieties was examined and the results are depicted in Fig. 6. The molecular descriptors like MW, hydrogen bond donor/acceptor, number of rotatable bonds, polar surface area, number of heavy atoms, and log*P* were computed for all the acyclic and cyclic non-redundant molecules. The distribution of descriptors is used to understand the property space of the molecules (Fig. 6), and thereby aid in estimating the druglikeness of the given chemical space.

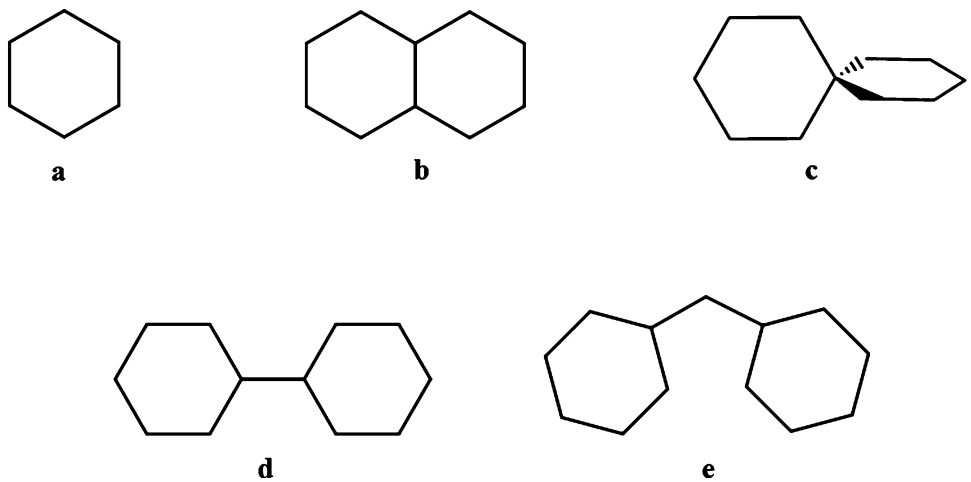


Fig. 5 A schematic representation of **a** unattached; **b** fused; **c** spiro; **d** connected; and **e** disjointed ring systems. The first four scaffolds represent unique features and therefore can correspond to a unique class. However, in case of **e** when a linker is involved, such an arrangement leads to a combination of more than one feature. This leads to the presence of more than one unique feature (correspond to a specific class) in a given molecule

Table 3 Pairwise distribution of 'n' membered rings as observed in MPDS-CL

	3	4	5	6	7	8
4	30,581					
5	641,651	404,334				
6	2,597,353	1,668,516	66,169,215			
7	22,867	15,344	312,595	1,699,025		
8	4931	2598	29,430	165,032	2013	
≥ 9	8460	2870	65,248	391,692	1587	1445

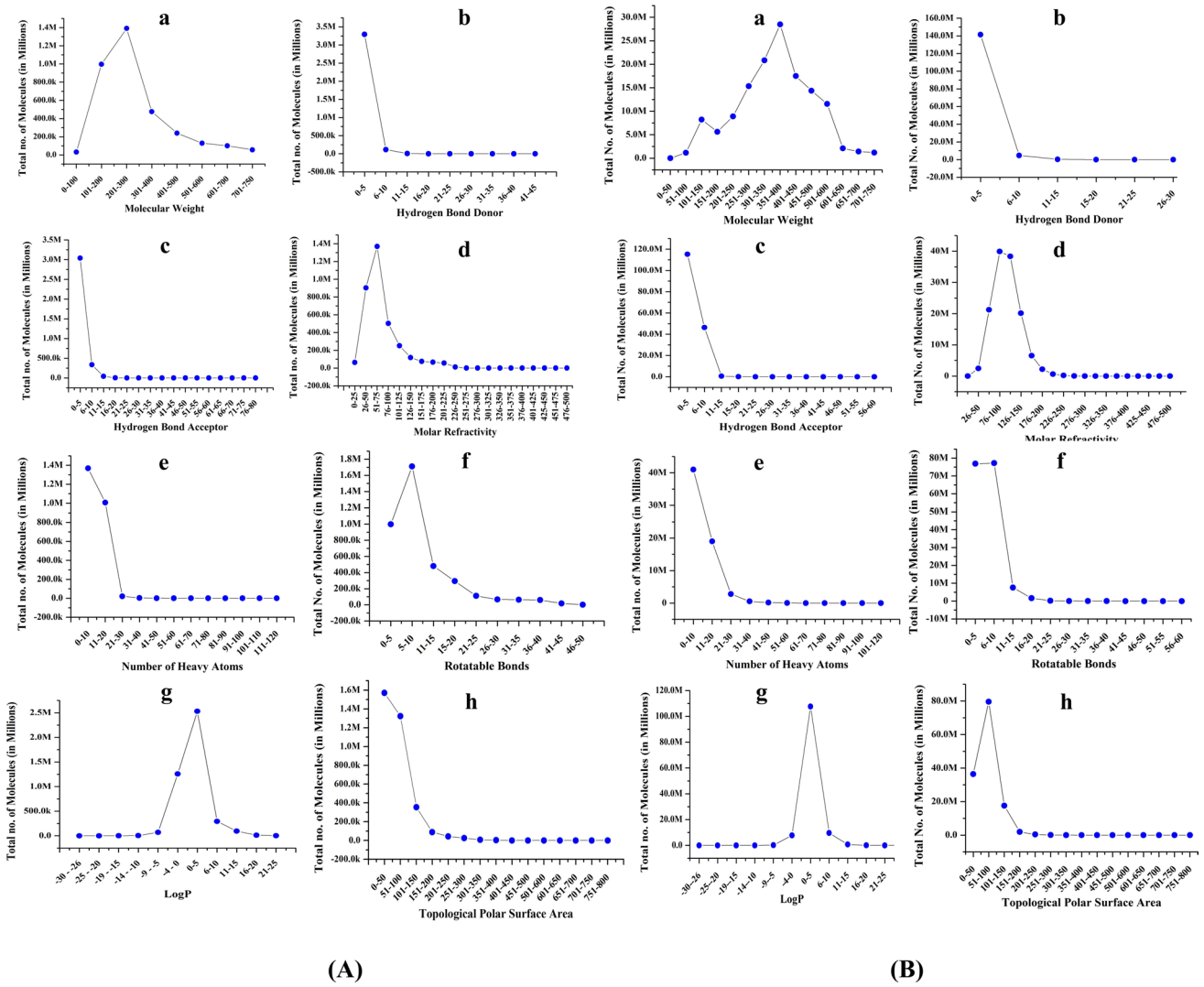


Fig. 6 Figure displaying the population-based distribution of selected molecular properties (for **A** acyclic and **B** cyclic molecules): (a) MW, (b) hydrogen bond donor, (c) hydrogen bond acceptor, (d) molar refractivity, (e) number of heavy atoms, (f) rotatable bonds, (g) logP and (h) topological polar surface area

Concerning MW, the highest number of molecules in acyclic space covers a range of 100.00–500.00 Da with a peak at 250.00 Da, while in the case of cyclic molecules, a progressive increase in the population of molecules is observed in the range of 150.00–750.00 Da, with the peak population between 300.00 and 350.00 Da. Figure 6 illustrates the distribution observed in, both cyclic and acyclic molecules with respect to their properties: (a) MW, (b) molar refractivity, (c) hydrogen bond donor, (d) hydrogen bond acceptor, (e) no. of heavy atoms, (f) rotatable bonds, (g) log*P*, and (h) topological polar surface area. The prototypical eight parameters considered for comparing and contrasting the chemical space distribution in cyclic and acyclic molecules reveal the following trends. Similar trends were observed in the distribution of hydrogen bond donor, hydrogen bond acceptor, logP and topological polar surface area for acyclic and cyclic molecules. While it has been observed that there is a slight broadening for cyclic molecules in the case of the distribution of MW, molar refractivity and the no. of heavy atoms. In contrast, in the case of rotatable bonds the distribution is sharper in cyclic molecules compared to acyclic. Thus, as expected the trends reveal higher diversity in the case of cyclic molecules and less flexibility compared to their acyclic counterparts.

Cheminformatics tools

Galaxy is a publicly available web server that provides an open-source web-based platform for a wide range of bioinformatics, cheminformatics, genomics, proteomics and

other analysis [29, 30, 51, 52]. Another major strength of Galaxy is the workflow system, which allows a very effective and automated execution of large projects involving multiple steps. Further, the ease with which one can write scripts, programs and develop software in any programming language to existing webservers based on Galaxy such as MPDS makes it highly desirable.

The chemical space in MPDS integrates various Galaxy chemical tools for structural analysis. Bray et al., developed the Galaxy ChemicalToolbox which provides the large assembly of tools for drug discovery and cheminformatics [31]. MPDS-CL provides access to the Galaxy ChemicalToolbox, as well as a host of other tools. These tools are PaDEL [53], CDK [54], Rdkit [49], Mordred [55], file format converters, JMol [56] editor and molecular visualizer for drawing and visualizing molecules along with a variety of search options (as described in Table 4 and Fig. 7). BCS classification, toxicity filter, drug likeness and natural product likeness filter are some other tools that are currently available and it is our quest to continually augment more filters like these. Efforts to add several of the in-house developed machine learning tools based on the properties of the molecule, such as the antiviral, toxicity, and susceptibility of failure in clinical trials, blood–brain barrier permeability prediction are in the pipeline [57–60].

The search methods and options

The search methods provided in the MPDS-CL is the way to navigate through the chemical space by employing various search options such as exact structure, sub-structure, fingerprint, and molecular property-based search. The idea of the exact structure search is to generate the exact molecule as provided by the user. Whereas the substructure search is employed to identify a series of molecules with the desired query. The substructure search is essentially employed to check for molecules that is built up with other scaffolds and in a way to explore the synergistic effects of different building blocks while interacting with the biological receptors. The results of the substructure search in MPDS-CL can be performed by providing a query molecule in.sdf/.mol/.smi format, which on search will result in giving the series of molecules with the substructure match in independently existing form. The fingerprint search is a search option to explore the molecules through specific classes. All classes are categorized under the main category viz., monocyclic, bicyclic, tricyclic, tetracyclic, and pentacyclic and a few special classes. The idea behind this search is to provide an understanding of the 56 classes, without asking the user to explicitly search for a specific molecule. Next in the line, is the search option based on molecular properties, where the user can search for molecules belonging to a range of properties as well as it comes with multiple filters to efficiently look for molecules with desired properties.

Table 4 Description of different modules available in MPDS–CL

Modules	Description
Get Data	Locally upload data/files of different file formats
Chemical structure editor	Draw a molecule (using Jmol editor) and export the SMILES
MPDS-AadharID search	A molecule from the MPDS-CL can be searched in various ways in addition to MPDS-AadharID-based search
Exact structure search	User can upload/draw structure and search
Sub-structure search	Search with sub-structure based on the user defined fragments
Properties-based search	Screening the molecules based on molecular properties
Fingerprint-based search	Identifying the structural features from the canonical SMILES
Fragmenter	Split a molecule to smaller fragments based on predefined rules (i.e., RECAP rules)
Fragment-based search	Searching the MPDS fragment library based on nature of fragments
File format conversion	Small molecules file format converter (i.e.,.pdb,.sdf,.smi,.inchi and.mol) for different cheminformatics operation
3D coordinates generation	Adding hydrogen atom to the molecule and convert the structure from 2 to 3D
Descriptors calculation	Calculation of different descriptors based on "PaDEL", "CDK", "RDkit", "Mordred"
Physico-chemical properties calculation	Calculating the physico-chemical properties for a set of molecules
Estimation of drug-likeness	Calculation of different drug-like rules, using DruLiTo, Lipinski's rule, Ghose filter, etc
BCS classification	Classifying the query molecule based on Biopharmaceutical Classification System (BCS)
Toxicity filter	Identifying the toxicophores for the given molecule
Natural product likeness calculator	Calculation of the natural product likeness score for the user given molecule

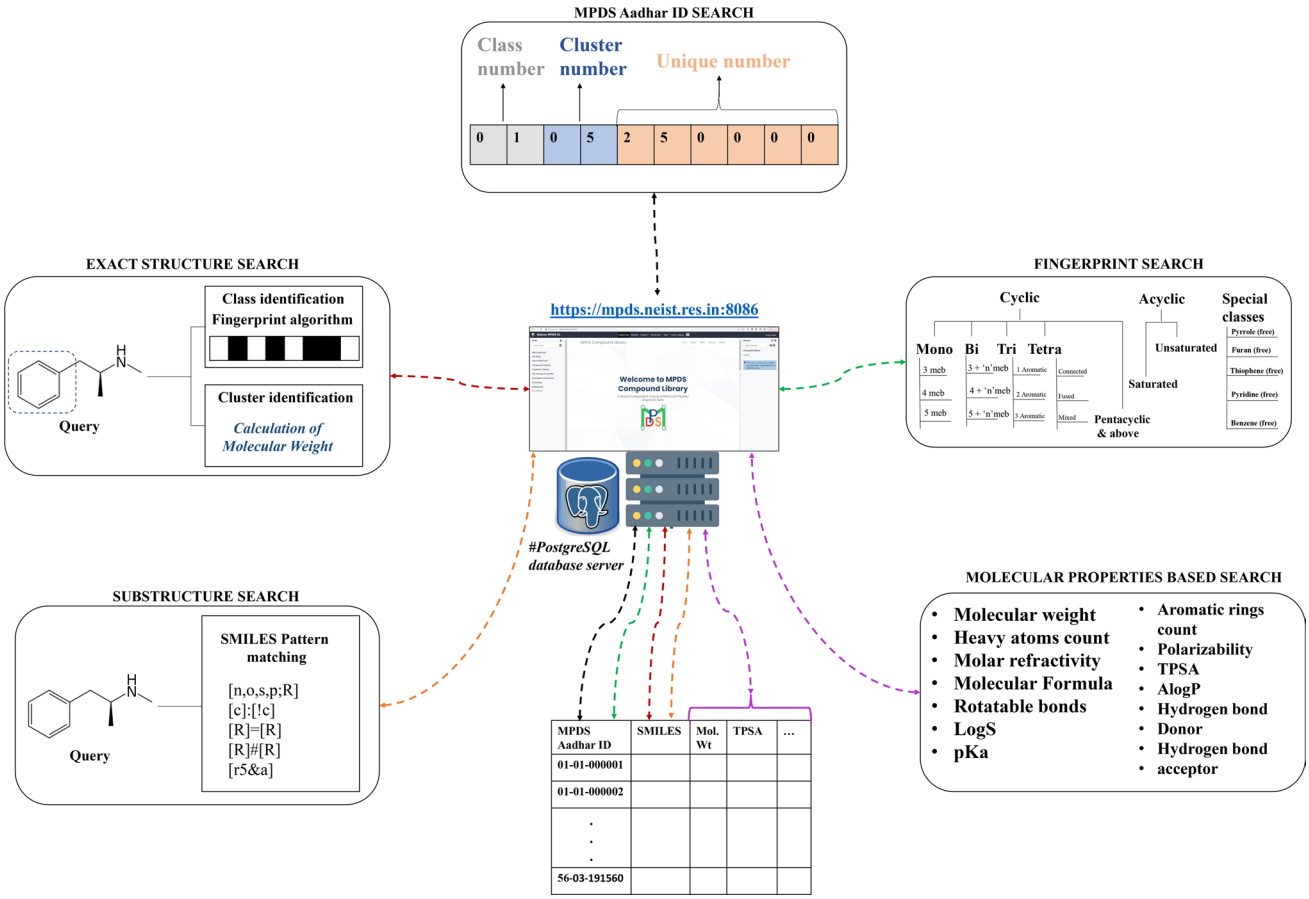


Fig. 7 A schematic diagram explaining different search methods available in MPDS-CL. The database information is connected with the PostgreSQL server. The query via MPDS-CL is connected to the database via Galaxy and fetches the information from the server

Each search method is designed to cater to a wide range of research objectives, allowing users to perform targeted searches as per their interests. The diverse search functionalities provided by the MPDS-CL enhance its utility as a valuable resource for researchers from various disciplines, empowering them to explore the chemical space, discover novel molecules, and gain valuable insights into compound properties (Fig. 7).

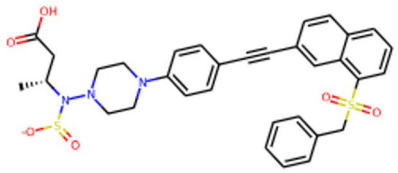
MPDS-AadharID

The name MPDS-AadharID is inspired by a thought process of comparing molecules to human beings. If a particular task has to be accomplished, naturally the focus is to find the right person, who is capable of executing the given task efficiently. Similarly, in the quest to discover a blockbuster drug, high value catalyst or any potentially useful molecule, the thrust will obviously be on to find the right molecule. As every molecule is unique in the chemical space, which is similar to a human in the population space, and therefore it has been our endeavour to assign a distinct ID to every molecule which traces its structural identity.

Each of these unique features has been clearly defined, and they have chosen to represent a "class" and thus, 56 classes were designated in MPDS-CL. As molecules can have multiple features, we employed priority rules and the class number is arranged in the ascending order. The 56-bit vector available in the MPDS-AadharID reveals the other features of a molecule, i.e., if a molecule belongs to class 45, but also contains the features of class 34, 23, and 8, then the 56-bit vector is 00000001000000000000001000000000010000000001000000000000.

The MPDS-AadharID of each existing molecule in the MPDS-CL is intended to provide all the information of a given molecule in a sequential fashion. The first page provides the critical details, which are common and computable to all the molecules in the chemical space, and it is generated on the fly by MPDS-CL (Fig. 8). However, molecules will have a varying detail of information and the subsequent pages of MPDS-AadharID can be custom designed to populate and use and as this information is specific to a given molecule and as such will not be generated on the fly.

MPDS AadharID No.: 32-19-122605



Molecular Formula:
$C_{33}H_{32}N_3O_6S_2$

IUPAC Name:
N-[(2R)-1-carboxypropan-2-yl]-4-{4-[2-(8-phenylmethanesulfonylnaphthalen-2-yl)ethynyl]phenyl}-N-sulfinatopiperazin-1-amine

Canonical SMILES:
S(=O)([O-])N(N1CCN(CC1)c1ccc(cc1)C#Cc1cc2c(S(=O)(=O)Cc3ccccc3)cccc2cc1)[C@@H](CC(=O)O)C

InChIKey:
UIEDIXWMFRQARL-RUZDIDTESA-M

Fingerprint:
00000000000010110001100000000001000000000000000000000000

Molecular Properties:

Mol. Wt.	630.754	LogP	3.25
HBD	1	LogS	-6.9
HBA	8	pKa	pKa1: 1.95; pKa2: 5.09; pKa3: 3.01; pKa4: -5.48
Molar refractivity	165.3	Polar surface area	121.29
Heavy atoms count	3,43,8,5	Aromatic Rings count	4
Rotatable bonds	10	Polarizability	67.17

Fig. 8 The first page of MPDS-AadharID generated from MPDS-AadharID search option of MPDS Compound Library. It depicts minimal critical information, which connects the unique MPDS-AadharID number with: **a** canonical SMILES, **b** InChIKey, **c** molecular formula, **d** 2D structure, **e** 56-bit fingerprint, **f** IUPAC name, and **g** molecular properties

Conclusions

MPDS-CL is a non-redundant chemical library represents about 150 million unique molecules, which was built by compiling a large dataset of molecules obtained from 42 publicly available chemical databases. It is an attempt to systematically classify the chemical space by infusing the structural-chemical insights which help in the design of molecules. The scheme of dividing molecules into 56 classes has been arrived at methodically by exploring various 'structural features' which determine the identity of a given molecule and all these molecules already available or easily accessible synthetically. The MPDS-CL provides various search options driven by MPDS-AadharID search, exact structure search, sub-structure search and fragment-based search, which helps in elegantly exploring the chemical space. The study has employed various cheminformatics, and other informatics methods to systematically analyse the chemical space and aid in the rational design of molecules with a desired property.

If one were to describe any effort to design a molecule, it is finding the right molecule for performing a given task. In the quest to discover a drug, catalyst or any special property of a molecule, it is all about hitting the right spot in the realm of chemical space. Further, understanding the structural and topological diversity of molecules and establishing various data-oriented analytics for structure–property and activity relationships is a topic of outstanding importance

in molecular design. Thus, when molecules are grouped in structurally similar categories, it unlocks newer possibilities for finding repurposable spectrums of varied interests and applications. Thus, the present work may be exploited in various fields, such as drug discovery, smart materials design, finding environmentally friendly pesticides, herbicides, petrochemicals, and other broad applications of chemical molecules.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11030-023-10752-1.

Acknowledgements 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).

Author contributions The entire project is conceived and designed by GNS. Bulk of the data collection and consolidation was done LJ, SN, HJM, NK, and AK. Analysis, validating the webserver, manual preparation was done by SV, AK, EJ, LP along with all others. The website design and validation are done by SV, HJM, SN, and LJ. First draft is prepared by LJ and LP. The manuscript was verified by all authors. The suggestion for the final draft was collected and a pre-final draft was made by LJ, LP, SN, and HJM. GNS has corrected and finalised the manuscript.

Data availability All the data can be obtained from the open-source platform provided in the article. The Python codes used to develop MPDS-CL is available on GitHub at https://github.com/gnsastry/MPDS-Compound-Library. These data are deposited in Zenodo, and it can be accessed at https://doi.org/10.5281/zenodo.8300413.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Reymond JL (2015) The chemical space project. Acc Chem Res 48(3):722–730. https://doi.org/10.1021/ar500432k
2. 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):384. https://doi.org/10.1038/s41597-020-00727-4
3. Warr WA, Nicklaus MC, Nicolaou CA, Rarey M (2022) Exploration of ultralarge compound collections for drug discovery. J Chem Inf Model 62(9):2021–2034. https://doi.org/10.1021/acs.jcim.2c00224
4. Koch MA, Schuffenhauer A, Scheck M, Wetzel S, Casaulta M, Odermatt A, Ertl P, Waldmann H (2005) Charting biologically relevant chemical space: a structural classification of natural products (SCONP). Proc Natl Acad Sci USA 102(48):17272–17277. https://doi.org/10.1073/pnas.0503647102
5. Pracht P, Bohle F, Grimme S (2020) Automated exploration of the low-energy chemical space with fast quantum chemical methods. Phys Chem Chem Phys 22(14):7169–7192. https://doi.org/10.1039/C9CP06869D
6. Nemoto S, Mizuno T, Kusuhara H (2023) Investigation of chemical structure recognition by encoder–decoder models in learning progress. J Cheminform 15(1):45. https://doi.org/10.1186/s13321-023-00713-z
7. Dunn TB, Seabra GM, Kim TD, Juárez-Mercado KE, Li C, Medina-Franco JL, Miranda-Quintana RA (2022) Diversity and chemical library networks of large data sets. J Chem Inf Model 62(9):2186–2201. https://doi.org/10.1021/acs.jcim.1c01013
8. Ertl P (2022) Magic rings: navigation in the ring chemical space guided by the bioactive rings. J Chem Inf Model 62(9):2164–2170. https://doi.org/10.1021/acs.jcim.1c00761
9. Flam-Shepherd D, Zhu K, Aspuru-Guzik A (2022) Language models can learn complex molecular distributions. Nat Commun 13(1):3293. https://doi.org/10.1038/s41467-022-30839-x
10. Asawa Y, Hatsuzawa S, Yoshimori A, Yamada K, Katoh A, Kouji H, Nakamura H (2021) Comprehensive exploration of chemical space using trisubstituted carboranes. Sci Rep 11(1):24101. https://doi.org/10.1038/s41598-021-03459-6
11. Vogt M (2023) Exploring chemical space—Generative models and their evaluation. Artif Intell Life Sci 3:100064. https://doi.org/10.1016/j.ailsci.2023.100064
12. Reymond JL, Van Deursen R, Blum LC, Ruddigkeit L (2010) Chemical space as a source for new drugs. Med Chem Commun 1(1):30–38. https://doi.org/10.1039/C0MD00020E
13. Coley CW (2021) Defining and exploring chemical spaces. Trends Chem 3:133–145. https://doi.org/10.1016/j.trechm.2020.11.004
14. Arve L, Voigt T, Waldmann H (2006) Charting biological and chemical space: PSSC and SCONP as guiding principles for the development of compound collections based on natural product scaffolds. QSAR Comb Sci 25(5–6):449–456. https://doi.org/10.1002/qsar.200540213
15. Gómez-Bombarelli R, Wei JN, Duvenaud D, Hernández-Lobato JM, Sánchez-Lengeling B, Sheberla D, Aguilera-Iparraguirre J, Hirzel TD, Adams RP, Aspuru-Guzik A (2018) Automatic chemical design using a data-driven continuous representation of molecules. ACS Cent Sci 4(2):268–276. https://doi.org/10.1021/acscentsci.7b00572
16. Sanchez-Lengeling B, Aspuru-Guzik A (2018) Inverse molecular design using machine learning: generative models for matter engineering. Science 361(6400):360–365. https://doi.org/10.1126/science.aat2663
17. Hoffmann T, Gastreich M (2019) The next level in chemical space navigation: going far beyond enumerable compound libraries. Drug Discov Today 24(5):1148–1156. https://doi.org/10.1016/j.drudis.2019.02.013
18. Kale B, Clyde A, Sun M, Ramanathan A, Stevens R, Papka ME (2023) ChemoGraph: interactive visual exploration of the chemical space. Comput Graph Forum 42(3):13–24. https://doi.org/10.1111/cgf.14807
19. Noguchi S, Inoue J (2022) Exploration of chemical space guided by PixelCNN for fragment-based de novo drug discovery. J Chem Inf Model 62(23):5988–6001. https://doi.org/10.1021/acs.jcim.2c01345
20. Rachman M, Piticchio S, Majewski M, Barril X (2021) Fragment-to-lead tailored in silico design. Drug Discov Today Technol 40:44–57. https://doi.org/10.1016/j.ddtec.2021.08.005
21. Gaur AS, John L, Kumar N, Vivek MR, Nagamani S, Mahanta HJ, Sastry GN (2023) Towards systematic exploration of chemical space: building the fragment library module in molecular property diagnostic suite. Mol Divers 27(3):1459–1468. https://doi.org/10.1007/s11030-022-10506-5
22. Bian Y, Xie XQ (2021) Generative chemistry: drug discovery with deep learning generative models. J Mol Model 27:71. https://doi.org/10.1007/s00894-021-04674-8
23. Gupta R, Srivastava D, Sahu M, Tiwari S, Ambasta RK, Kumar P (2021) Artificial intelligence to deep learning: machine intelligence approach for drug discovery. Mol Divers 25:1315–1360. https://doi.org/10.1007/s11030-021-10217-3

24. Karthikeyan A, Priyakumar UD (2022) Artificial intelligence: machine learning for chemical sciences. J Chem Sci 134:1–20. https://doi.org/10.1007/s12039-021-01995-2
25. Murugan NA, Priya GR, Sastry GN, Markidis S (2022) Artificial intelligence in virtual screening: models versus experiments. Drug Discov Today 27(7):1913–1923. https://doi.org/10.1016/j.drudis.2022.05.013
26. von Lilienfeld OA, Burke K (2020) Retrospective on a decade of machine learning for chemical discovery. Nat Commun 11(1):4895. https://doi.org/10.1038/s41467-020-18556-9
27. Wigh DS, Goodman JM, Lapkin AA (2022) A review of molecular representation in the age of machine learning. Wiley Interdiscip Rev Comput Mol Sci 12(5):e1603. https://doi.org/10.1002/wcms.1603
28. Djoumbou Feunang Y, Eisner R, Knox C, Chepelev L, Hastings J, Owen G, Fahy E, Steinbeck C, Subramanian S, Bolton E, Greiner R (2016) ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. J Cheminform 8(61):1–20. https://doi.org/10.1186/s13321-016-0174-y
29. Galaxy Community (2022) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. Nucleic Acids Res. 50(W1):W345–W351. https://doi.org/10.1093/nar/gkac247
30. Gu Q, Kumar A, Bray S, Creason A, Khanteymoori A, Jalili V, Grüning B, Goecks J (2021) Galaxy-ML: an accessible, reproducible, and scalable machine learning toolkit for biomedicine. PLOS Comput Biol 17(6):e1009014. https://doi.org/10.1371/journal.pcbi.1009014
31. Bray SA, Lucas X, Kumar A, Grüning BA (2020) The ChemicalToolbox: reproducible, user-friendly cheminformatics analysis on the Galaxy platform. J Cheminform 12(1):1–7. https://doi.org/10.1186/s13321-020-00442-7
32. 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 129:515–531. https://doi.org/10.1007/s12039-017-1268-4
33. 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 28(11):913–926. https://doi.org/10.1080/1062936X.2017.1402819
34. 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 Inform 85:114–125. https://doi.org/10.1016/j.jbi.2018.08.003
35. Gaur AS, Nagamani S, Priyadarsinee L, Mahanta HJ, Parthasarathi R, Sastry GN (2023) Galaxy for open-source computational drug discovery solutions. Expert Opin Drug Discov 18(6):579–590. https://doi.org/10.1080/17460441.2023.2205122
36. Xie Y, Xu Z, Ma J, Mei Q (2022) How much of the chemical space has been explored? selecting the right exploration measure for drug discovery. In: Proceedings of ICML 2022 2nd AI for science workshop
37. van Vlijmen H, Ortholand JY, Li VM, de Vlieger JSB (2021) The European Lead Factory: an updated HTS compound library for innovative drug discovery. Drug Discov Today 26(10):2406–2413. https://doi.org/10.1016/j.drudis.2021.04.019
38. Medina-Franco JL, Martínez-Mayorga K, Giulianotti MA, Houghten RA, Pinilla C (2008) Visualization of the chemical space in drug discovery. Curr Comput Aided Drug Des 4(4):322–333. https://doi.org/10.2174/157340908786786010
39. 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
40. Reddy AS, Pati SP, Kumar PP, Pradeep HN, Sastry GN (2007) Virtual screening in drug discovery—a computational perspective. Curr Protein Pept Sci 8(4):329–351. https://doi.org/10.2174/138920307781369427
41. Priyadarsinee L, Jamir E, Nagamani S, Mahanta HJ, Kumar N, John L, Sarma H, Kumar A, Gaur AS, Sahoo R, Vaikundamani S, Murugan NA, Priyakumar UD, Raghava GPS, Bharatam PV, Parthasarathi R, Subramanian V, Sastry GM, Sastry GN (2023) Molecular property diagnostic suite for COVID-19 (MPDS-COVID-19): an open access disease specific drug discovery portal. bioRxiv. https://doi.org/10.1101/2023.08.29.555437
42. Druzhilovskiy DS, Rudik AV, Filimonov DA, Gloriozova TA, Lagunin AA, Dmitriev AV, Pogodin PV, Dubovskaya VI, Ivanov SM, Tarasova OA, Bezhentsev VM, Murtazalieva KA, Semin MI, Maiorov IS, Gaur AS, Sastry GN, Poroikov VV (2017) Computational platform Way2Drug: from the prediction of biological activity to drug repurposing. Russ Chem Bull 66:1832–1841. https://doi.org/10.1007/s11172-017-1954-x
43. Murtazalieva KA, Druzhilovskiy DS, Goel RK, Sastry GN, Poroikov VV (2017) How good are publicly available web services that predict bioactivity profiles for drug repurposing? SAR QSAR Environ Res 28(10):843–862. https://doi.org/10.1080/1062936X.2017.1399448
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):1–4. https://doi.org/10.1186/1758-2946-3-33
45. Kiewhuo K, Gogoi D, Mahanta HJ, Rawal RK, Das D, Sastry GN (2022) North East India medicinal plants database (NEI-MPDB). Comput Biol Chem 100:107728. https://doi.org/10.1016/j.compbiolchem.2022.107728
46. Kiewhuo K, Gogoi D, Mahanta HJ, Rawal RK, Das D, Vaikundamani S, Jamir E, Sastry GN (2023) OSADHI - An online structural and analytics-based database for herbs of India. Comput Biol Chem 102:107799. https://doi.org/10.1016/j.compbiolchem.2022.107799
47. Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D (2015) InChI, the IUPAC International Chemical Identifier. J Cheminform 7(1):1–34. https://doi.org/10.1186/s13321-015-0068-4
48. Van Rossum G, Drake FL (1995) Python reference manual, vol 111. Centrum voor Wiskunde en Informatica, Amsterdam, pp 1–52
49. Landrum G (2013) Rdkit documentation. Release 1(4):1–79
50. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3(1):1–9. https://doi.org/10.1038/sdata.2016.18

51. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W (2005) Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15(10):1451–1455. https://doi.org/10.1101/gr.4086505
52. Goecks J, Nekrutenko A, Taylor J, Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11:1–13. https://doi.org/10.1186/gb-2010-11-8-r86
53. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 32(7):1466–1474. https://doi.org/10.1002/jcc.21707
54. Willighagen EL, Mayfield JW, Alvarsson J, Berg A, Carlsson L, Jeliazkova N, Kuhn S, Pluskal T, Rojas-Chertó M, Spjuth O, Torrance G (2017) The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. J Cheminform 9:1–9. https://doi.org/10.1186/s13321-017-0220-4
55. Moriwaki H, Tian YS, Kawashita N, Takagi T (2018) Mordred: a molecular descriptor calculator. J Cheminform 10:1–4. https://doi.org/10.1186/s13321-018-0258-y
56. Hanson RM (2010) Jmol—a paradigm shift in crystallographic visualization. J Appl Crystallogr 43(5):1250–1260. https://doi.org/10.1107/S0021889810030256
57. John L, Soujanya Y, Mahanta HJ, Sastry GN (2022) Chemoinformatics and machine learning approaches for identifying antiviral compounds. Mol Inform 41(4):e2100190. https://doi.org/10.1002/minf.202100190
58. John L, Mahanta HJ, Soujanya Y, Sastry GN (2023) Assessing machine learning approaches for predicting failures of investigational drug candidates during clinical trials. Comput Biol Med 153:106494. https://doi.org/10.1016/j.compbiomed.2022.106494
59. Mazumdar B, Sarma PKD, Mahanta HJ, Sastry GN (2023) Machine learning based dynamic consensus model for predicting blood-brain barrier permeability. Comput Biol Med 160:106984. https://doi.org/10.1016/j.compbiomed.2023.106984
60. 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 138:104856. https://doi.org/10.1016/j.compbiomed.2021.104856
61. Kim S et al (2023) PubChem 2023 update. Nucleic Acids Res 51:D1373–D1380
62. Mcule Database. https://mcule.com/database/. Accessed 18 Jul 2023
63. eMolecule Database. https://www.emolecules.com/info/plus/download-database/. Accessed 18 Jul 2023
64. Papadatos G et al (2016) SureChEMBL: a large-scale, chemically annotated patent document database. Nucleic Acids Res. 44(D1):D1220–D1228
65. Henry VJ et al (2014) OMICtools: an informative directory for multi-omic data analysis. Database. https://omictools.com/cococotool/. Accessed 18 Jul 2023
66. Gaulton, A., et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.*, 40(D1), D1100-D1107. https://www.ebi.ac.uk/chembl/. Accessed 18 Jul 2023
67. Chemdiv Database. https://www.chemdiv.com/. Accessed 18 Jul 2023
68. SPECS Database. https://www.specs.net/. Accessed 18 Jul 2023
69. Gaulton A et al (2017) The ChEMBL database in 2017. Nucleic Acids Res. 45(D1):D945–D954
70. Ligand Info Database. http://Ligand.Info. Accessed 18 Jul 2023
71. GPCR Decoy Database. https://cavasotto-lab.net/Databases/GDD/. Accessed 18 Jul 2023
72. Liu T et al (2007) BindingDB: a web-accessible database of experimentally determined protein–ligand binding affinities. Nucleic Acids Res. 35(suppl_1):D198–D201
73. TimTec Database. https://www.timtec.net/. Accessed 18 Jul 2023
74. Asinex Database. https://www.asinex.com/. Accessed 18 Jul 2023
75. InterBioScreen. https://www.ibscreen.com/. Accessed 18 Jul 2023
76. Sorokina M et al (2021) COCONUT online: collection of open natural products database. J. Cheminform 13(1):1–13
77. Universal Natural Products Database. https://unaprod.com/. Accessed 18 Jul 2023
78. NCI. https://cactus.nci.nih.gov/. Accessed 18 Jul 2023
79. Crystallography Open DB Database. http://www.crystallography.net/cod/. Accessed 18 Jul 2023
80. Wishart DS et al (2022) HMDB 5.0: the human metabolome database for 2022. Nucleic Acids Res. 50(D1):D622–D631
81. Skuta C et al (2017) Probes & drugs portal: an interactive, open data resource for chemical biology. Nat Methods 14(8):759–760
82. Openmolecules Database. https://openmolecules.org/. Accessed 18 Jul 2023
83. Lipid Bank Database. http://lipidbank.jp/. Accessed 18 Jul 2023
84. ChemBank Database. https://data.broadinstitute.org/chembank/. Accessed 18 Jul 2023
85. Backman TWH et al (2011) ChemMine tools: an online service for analyzing and clustering small molecules. Nucleic Acids Res. 39(suppl_2):W486–W491
86. ChEBI Database. https://www.ebi.ac.uk/chebi/. Accessed 18 July 2023
87. Frolkis A et al (2010) SMPDB: the small molecule pathway database. Nucleic Acids Res. 38(suppl_1):D480–D487
88. GPCR Ligand Database. https://gpcrdb.org/. Accessed 18 Jul 2023
89. Kanehisa M (2002) The KEGG database. In: 'In silico' simulation of biological processes: Novartis Foundation Symposium, vol 247. Wiley, Chichester, pp 91–103. https://www.genome.jp/kegg/drug/. Accessed 18 Jul 2023
90. MDPI. https://www.mdpi.org/cumbase.htm. Accessed 18 Jul 2023
91. TOSLab. https://toslab.no/Hje. Accessed 18 Jul 2023
92. Wishart DS et al (2018) DrugBank 50: a major update to the DrugBank database for 2018. Nucleic Acids Res. 46(D1):D1074–D1082
93. MyriaScreen. http://www.myriascreen.com/. Accessed 18 Jul 2023
94. GRAC. https://www.guidetopharmacology.org/. Accessed 18 Jul 2023
95. PHARMGKB. https://www.pharmgkb.org/. Accessed 18 Jul 2023
96. ChemDB. https://cdb.ics.uci.edu/. Accessed 18 Jul 2023
97. Ursu O et al (2016) DrugCentral: online drug compendium. Nucleic Acids Res 45(D1):D932–D939
98. SuperDrug2. http://bioinf.charite.de/superdrug/. Accessed 18 Jul 2023
99. Therapeutic Target Database. https://db.idrblab.net/ttd/. Accessed 18 Jul 2023
100. Irwin JJ, Shoichet BK (2005) ZINC—a free database of commercially available compounds for virtual screening. J Chem Inf Model 45(1):177–182