Medicinal chemistry lens on patent data in SureChEMBL

Yojana Gadiya[1,2,3,†,*], Simran Shetty[1,2,4,†], Martin Hofmann-Apitius[3,5], Philip Gribbon[1,2,], and Andrea Zaliani[1,2]

1. Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Schnackenburgallee 114, 22525 Hamburg, Germany
2. Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), Theodor Stern Kai 7, 60590 Frankfurt, Germany
3. Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53113 Bonn, Germany
4. Hamburg University of Applied Sciences (HAW), 20099 Hamburg, Germany
5. Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven 53757 Sankt Augustin, Germany

† These authors contributed equally to this work.

*** Corresponding author:** Yojana Gadiya, Fraunhofer Institute for Translational Medicine and Pharmacology
Email: Yojana.Gadiya@itmp.fraunhofer.de

Abstract

In the pharmaceutical industry, the patent protection of drugs and medicines is accorded particular importance because of the high costs of research and development (R&D), associated with the considerable risks inherent in the creation of novel drugs. Consequently, patents play a pivotal role in ensuring a suitable return on R&D investments, especially those directed toward clinical trials. The patent corpus has been analyzed by researchers to identify freedom-to-operate spaces for novel drug candidates. Several well-established public patent data repositories have enabled automated methodologies for extracting information on therapeutic agents. In this manuscript, we delve into one such publicly available patent database, SureChEMBL, which catalogues patents related to life sciences. Our exploration begins by evaluating the coverage of compounds within SureChEMBL in comparison to other chemical data resources. Additionally, we pinpoint critical sections of patent documents where detailed chemical annotations can be found. Next, we exhibit the compounds' potential to serve as drug candidates by evaluating their conformity to the Rule of Five (Ro5) and its extension, the beyond Ro5 (bRo5). Furthermore, we examine the development stage (preclinical or clinical) reported for these compounds as documented in public databases like ChEMBL. In summation, our investigation aims to provide an estimate of the relevance of the compounds catalogued within the SureChEMBL database in the context of the drug discovery domain.

Keywords: Patent, FAIR, SureChEMBL, Clinical-patent exploration, Ro5 bRo5 categorization

1. Introduction

Patents are legal documents that disclose an invention to the public (https://www.wipo.int/patents/en/). With this disclosure, the holder of a valid patent generally has the exclusive right to make, use, and sell the invention for approximately 20 years [1, 2]. In drug discovery, drug hunters extensively published patents, to identify competing interests among pharmaceutical companies and research institutes regarding potential drug candidates and marketed compounds[3]. Additionally, patent documents serve as a catalyst for medicinal chemists, empowering them to optimize their drug candidates strategically and ensure their alignment with freedom-to-operate (FTO) zones that may exist outside of the scope of the claimed patent coverage [4].

Pharmaceutical-based patenting activity, which mainly covers claims related to a therapeutic design, synthesis, formulation, processing, and alternative usage or indications, reveals critical information pertaining to the development and prescription of drugs and biologics. In doing so, it serves as a valuable resource for understanding the landscape and dynamics of the pharmaceutical industry [5-9]. Pharmaceutical patents cover

two fundamental components: the compound itself and its application [10, 11]. The compound is usually identified in its various forms, such as within a Markush structure, a trade name, a generic name, a lab code, and/or a CAS Registry number. Patents claim a compound by its structure or even claim a family of structures (based on a scaffold). The chemical structure information is the basis for conducting chemical patent searches by scientists, chemistry patent attorneys, health professionals, chemists, and professionals in the pharmaceutical industry and is leveraged by commercial vendors in the form of expert software tools and services (eg. CAS-SciFinder) [12]. The application field(s) of a patent is usually found in the claims or description sections of the patent in text format. A claim's text description is a legally focused document that often contains specialized terminology and jargon integral to the patent domain. This content plays a crucial role in defining the scope of the underlying patent, making it a subject of study for patent lawyers and pharma R&D scientists who seek to comprehend the FTO space associated with the patents.

Pharmaceutical patent analysis is a well-established domain in the literature covering a wide range of topics, including patenting trends [13], tools for patent protection [14], patent portfolio analysis [15], identification of novel chemical entities for certain treatment or diseases [16], identification of the key players [17], and many more. In the study by Falaguera and Mestres (2021), for instance, compounds mined from patents and deposited in the public patent database (especially in SureChEMBL) were found to consist of a collection of either starting materials, intermediate products, or pharmacologically relevant compounds [18]. Of these, only the pharmacologically relevant compounds play a significant role in drug discovery, and hence chemoinformatic techniques, such as the matched molecular pair (MMP) [19] and maximum common substructures (MCS) [18], became prominent alternatives to the generic Markush structural searches [20, 21] to identify these compounds. Additionally, methods for structural elucidation of pharmacologically relevant compounds in patent documents have allowed for the generation of metrics that assess the innovation involved in the development of drugs, as studied by Wills and Lipkus (2020) [22]. Despite all these efforts, the majority of the aforementioned analyses have been limited to patent documents filed and/or granted in the United States of America (USA). This is attributed to three main reasons: i) the USA is the world's largest technology market [23] in general and the pharmaceutical market, in particular, ii) the presence of easy-to-use US-centric public patent databases, such as those managed by United States Patent and Trademark Office (USPTO), which allow for bulk download of patent documents and their metadata [24], and iii) the availability of resources, such as the FDA's Orange Book, allowing for tracking of drug candidates and their corresponding patents through time [25]. Furthermore, these analyses confine the characterization of pharmaceutical patents to those specifically labeled with the IPC code A61K and it is known that this IPC class includes hygiene-related patent documents, in addition to medicinal ones, possibly merging not pharma relevant annotations.

In this manuscript, we aim to understand the relevance of compounds annotated by SureChEMBL's pipeline and understand their correlation to drugs available in the market. Rather than limiting to patents found in the USA, as done in all the existing analyses so far, we elucidate a diverse dataset of patent applications filed and granted all over the world (US; 50.8%, World; 28.5%, and Europe; 20.6%). Moreover, we covered larger IPC patent classes that include information on the medicinal utility of compounds and the formulation of compounds, thus covering a wider chemobiology domain. Our exploration scrutinises the availability of patent-described compounds beyond their patent documents, specifically those annotated by SureChEMBL and other public compound databases. Subsequently, we examined the innovation, preferences in medicinal chemistry and indications of procedural steps embraced by pharmaceutical companies to facilitate medicine discovery. We conclude by delving into the clinical candidate space of the patented compounds in order to understand the success rate of progressing compounds from patent application to clinical practice.

2. Methods

2.1. Collecting compound data from patent literature

We used SureChEMBL (https://www.surechembl.org/), an extensive publicly available patent chemical data catalogue, as a source for patent documents and metadata [26]. The database extracts compounds from patent documents by automated text and image analysis pipelines. Moreover, each extracted patent compound is

associated with its structural information (i.e., SMILES, InChIKey, InChI) and the textual location of extraction in the patent document, such as claims, title, description, etc. We used the compounds patented from 2015-2022 as the patent chemical collection in the manuscript. The legacy data (from 1994-2014) had a different data format with the absence of patent information making it difficult to transition from the compounds to the corresponding patents. As a result of this, the legacy data was omitted for the analysis pertaining to the manuscript.

2.2. Compound databases utilized for compound metadata annotation

To normalize the chemical structures in the resource, we mapped them to three large independent compound data resources, namely PubChem (version 2023) [27], ChEMBL (version 32) [28], and DrugBank (version 5.1.10) [29]. PubChem (https://pubchem.ncbi.nlm.nih.gov/), maintained by the National Institutes of Health (NIH), contains information including the chemical structure of compounds, associated bioactivities, patents, and quality control data. PubChem's European equivalent, ChEMBL (https://www.ebi.ac.uk/chembl/), is maintained by the European Bioinformatics Institute. It is an SQL data resource connecting compounds to bioactivity, patents, and disease mechanisms. On the other hand, DrugBank (https://go.drugbank.com/) focuses on clinically approved drug candidates, including information on pharmacodynamics and pharmacokinetic properties.

The ChEMBL "MaxPhase" field was used to annotate the clinical stage of compounds, based upon categories shown in **Table 1**. The Drugbank resource was used to validate the compounds having reached the stage of currently approved drugs and distinguish them from those that had been withdrawn, thus providing insights into the compound's potential applicability and long-term effects. The commercial nature of Drugbank, while assuring faster updates than public databases, might hinder certain information from being available only to premium subscribers.

Clinical stage	Description
Preclinical	Compounds that have bioactivity data but no clinical data available.
Phase 1	Compounds currently registered or are part of Phase 1 trials or have passed the Phase 1 trials and are being registered for Phase 2.
Phase 2	Compounds that have passed the Phase 2 trials and are being registered for Phase 3.
Phase 3	Compounds that are currently in Phase 3 trials.
Approved	Compounds that have been approved for certain indications.
Unknown	Compounds that could not be mapped to any clinical trial information for a specific indication.

Table 1: ChEMBL clinical or research stage annotation definitions for compounds.

2.3. Chemical space exploration using physicochemical properties of compounds

To improve the efficiency of the drug discovery process, scientists have formulated guidelines on key determinants of drug likeness. As drug discovery pivoted towards high throughput screening in the mid-1990s, emphasis was placed on generating large chemical libraries and selecting potential candidates from these libraries [30]. Lipinski's Rule of Five (Ro5) was the first set of rules to be developed in 1997 (modified in 2000) in this context [31, 32] These rules stated that a (non-antimicrobial) compound is more likely to reach the clinic as an oral drug if following conditions are satisfied:

- molecular weight (MW) < 500 Daltons
- showed lipophilicity (LogP) < 5
- contained < 5 hydrogen bond donor atoms (HBD),
- and contained < 10 hydrogen-bond acceptor atoms (HBA)

Along with the Ro5 rules, Veber *et al.* (2002) added rotatable bonds (NRotB) and the topological polar surface area (TPSA) as important features, highlighting the compounds should have a TPSA < 140 and NRotB < 12 to enhance the probability of sufficient oral bioavailability [33]. Hence, to profile the drug-like space for the patented compounds, we compared them using Veber's rules such that all conditions must be satisfied for it to be considered orally bioactive.

In follow-on analyses, Doak *et al.* (2014), with other researchers, updated the Ro5 for oral bioavailability space of drugs (initially described Ro5 described by Lipinski *et al.,* 2000) and named these drug-like spaces as the beyond Ro5 (bRo5) space [34-36]. The criteria of bRo5 supported the selection of cell-permeable clinical candidates showing good pharmacokinetics (PK) and explored the "undruggable" targets, which could not have been possible previously with the Ro5 filtering. We adopt the definition of the bRo5 space from that of Doak *et al.* (2016), considering compounds with MW > 500 daltons with at least one property beyond the Ro5 (LogP > 7, HBD > 5, HBA > 10, TPSA > 200, and NRotB > 20) [37].

Scaffold diversity analysis is a widely used method for assessing the diversity of compounds, focusing on their structural and activity-related aspects. By employing scaffold diversity assessments, medicinal chemists can broaden their knowledge of chemical space by introducing novel compounds. In our exploration of chemical space within patent documents, we conducted a scaffold diversity assessment on compounds derived from these patents. To accomplish this, we simplified the desalted compounds into their Murcko scaffolds, preserving the generic forms of ring components, linkers, and side chains [38]. These Murcko scaffolds were then used to examine the occurrence of newly registered scaffolds on an annual basis.

Pan-assay interference compounds (PAINS) are frequently used in drug discovery to flag and mark compounds that would cause interference in bioassays due to their non-specific polypharmacological effects [39]. The compounds showing these substructural alters are usually dropped off at the earlier stages of the drug discovery pipeline due to their low chances of success through the clinical pipeline. To explore whether compounds that are reactive, toxic, or chemically labile groups are part of patent documents, we reviewed the compounds for PAINS filter alters.

2.4. Implementation details

Analyses were conducted in Python (version 3.10) with mapping of chemicals to database identifiers using RDKit [40] and PubChempy [41]. The PAINS filtering was performed using KNIME's RDKit molecular filter node. The desalting of compounds was performed by the Vernalis desalting node in KNIME.

2.5. Data availability

All the Python scripts and corresponding data files used in this manuscript are available on GitHub (https://github.com/Fraunhofer-ITMP/patent-clinical-candidate-characteristics) and Zenodo (https://doi.org/10.5281/zenodo.10210061) respectively.

3. Results

Our dataset included a collection of 10 million (10,718,652 based on unique SureChEMBL IDs and 10,686,744 based on unique InChIKeys) compounds found in over 1 million (1,574,602) patent applications (including both granted and non-granted) between 2015 and 2022. We refer to patent compounds as those that were captured and annotated by SureChEMBL's internal pipeline as compounds associated with patents. On average, we found that 196,826 patent applications and about 1M (1,137,183) compounds have been filed and patented each year within the life science sector respectively **(Figure 1)**. Moreover, an average ratio of 6 compounds registered per patent was found.

Of these million patented compounds, we found that 0.2% of compounds were associated with multiple patent documents that would involved repurposing patents (i.e., patents talking about reusing known drugs for a different indication) or compounds annotated from utility patents (i.e., a patent application dealing with approved drugs in different pharmaceutical forms or route of administration for specific treatments) **(Supplementary Figure S1)**. Moreover, the distribution of the patents associated with each compound showed that 95% of compounds were found in less than 5 patent documents **(Supplementary Table T1)**. Lastly, we explored the patent applications' country and type (filed vs. granted). Based on the country associated with the patents, the United States (US) and Europe (EP) were predominant, with 57.3% (with 24% granted) and 26.6% (with 11% granted) patents, respectively. SureChEMBL claims to aggregate compounds from Japan (JP) as well, but the contribution of these patents in our dataset was found to be less than 1% [26]. This could be due to the complexity involved in the conversion of Japanese text to English, as a result of which SureChEMBL restricts the ingestion and storing\ of only English-based abstracts [44]. Additionally, the patent numbering and coding systems for Japanese patents differ significantly, which may make it more difficult to search for and extract information from Japanese patents [42]. Also, legacy data (before 2015) might involve a larger Japanese cohort as claimed by SureChEMBL which was not part of this study.

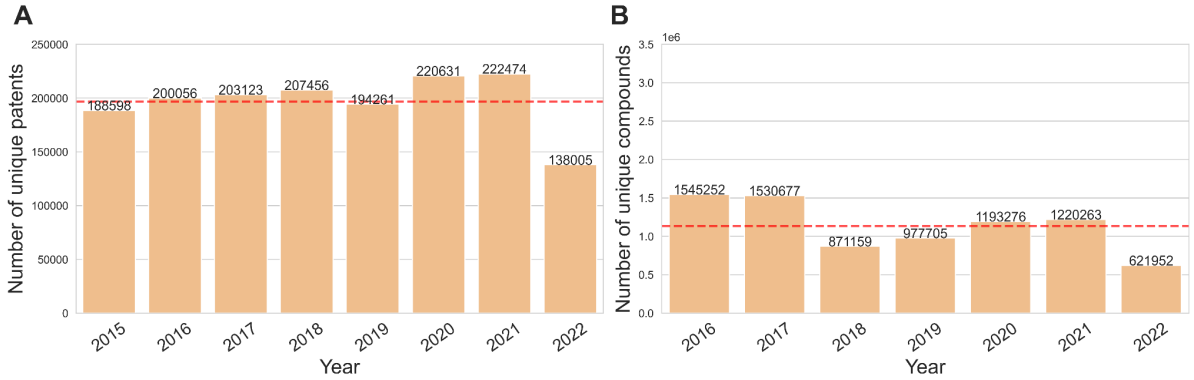


Figure 1: Data distribution over the years. A) Annual unique patents filed over the years. B) Annual unique compounds found within patent documents. The red line in both plots indicates the baseline with the average counts of the respective entities.

In the following sections, we will explore the compound and patent landscape of the aforementioned data. We start by discussing the searchability of the patented compounds within public databases and the findability of the compounds within patent documents. In the context of this manuscript, searchability is defined as the ability to search for the compounds found in patent documents using their molecular representations (i.e. SMILES or InChiKey) in databases, whereas findability is defined as the ability to trace the textual location of the compound in the patent document. Next, we look into the structural diversity of patented compounds and their drug-likeness discussing the prospective proportion of compounds in the database that can be reused or repurposed. Moreover, we look into the underlying chemical space to enumerate and identify patterns observed within the patented compounds. In the end, we discuss in brief the success rate of a compound from patent to clinical candidate from the SureChEMBL compound space.

3.1. Cross-resource comparison of compounds in patents

A large number of compounds-centric biology databases have been established in the past decades [43-46]. These databases have served various purposes in drug discovery, from identifying the bioactivity of unknown compounds [47, 48] to the prediction of mechanism-of-action (MoA) [49, 50] or simply for the virtual screening of drug candidates [51, 52]. While previous research by Joerg Ohms (2021) explored the coverage of patent documents in relation to chemicals across patent databases, the author was restricted to comparing the compound coverage between PubChem and SureChEMBL [53]. Thus, to better understand the coverage of patent compounds in prominent chemical databases (PubChem, ChEMBL, and DrugBank), we analyzed the structural overlap between the compounds cited in patents and those found in the three public databases. This comparison revealed two key findings discussed below.

In the first, only 0.02% (2,096 out of 10 million) of the compounds found within SureCHEMBL patent space had been approved for one or more indication areas, according to DrugBank **(Figure 2)**. This low percentage could be due to three reasons: *i)* crafting of patent documents that intentionally encompass a wider chemical structure-activity landscape than the clinical candidates, thereby safeguarding the latter and their secretness [54]; *ii)* it has been pointed out that the nomenclature of the drug candidate changes as the drug candidate progresses through the clinical pipelines of drug discovery and hence could also be a reason for not being able to identify the compounds in the database as we will discuss in the later sections; and *iii)* DrugBank is a commercial database, and limited information is provided for academic research, which could also be a factor responsible for the low overlap of approved drugs across patent documents.

The second finding was that PubChem-retrieved compounds exhibited the highest overlap (94.4%) with the SureCHEMBL patented compounds **(Figure 2)**. This is ascribed to the extensive coverage of PubChem, encompassing 115 million compounds from 925 resources, and its automated patent annotation pipeline that established connections between the compounds in the database to relevant patent documents referenced in Google Patents in its latest updates [27]. On the other hand, ChEMBL demonstrates only a 3% overlap with compounds from patent applications, a sign that ChEMBL and SureChEMBL differ heavily in chemical space as the patent compounds are rarely uploaded with any biology activity (assays) into ChEMBL by authors. Furthermore, it also indicates that the compounds not found in the public data resources (ChEMBL, PubChem, and DrugBank) are potentially restricted to commercial compound libraries used by patent assignees for drug discovery.

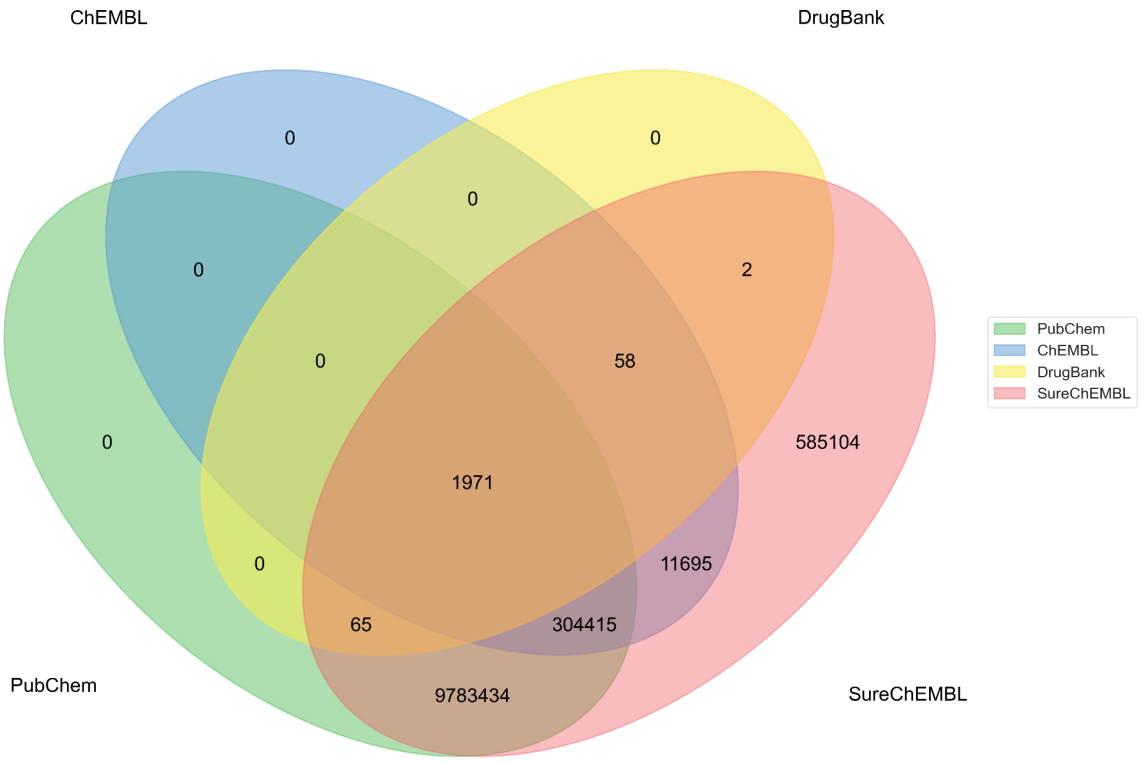


Figure 2: Distribution of patented compounds across four chemical resources, namely SureCHEMBL, PubChem, ChEMBL, and DrugBank data resources. The largest overlap of SureChEMBL was found with PubChem, followed by ChEMBL, and the least in DrugBank. Moreover, most compounds within DrugBank were found to already exist in ChEMBL and PubChem.

3.2 Discovering chemicals in patents by assessing their findability

A patent is a structured document containing pre-defined sections such as the title, abstract, description, and claims [55, 56]. Within these documents are sections where the context in which a compound's information is displayed could provide hints about its applicability in the patent. For instance, if a compound is mentioned in the background section it is likely to be associated with prior art relevant to the patent, whereas a compound mentioned in the summary or description sections would pertain to the novel invention disclosed in the patent document. Hence, our investigation focused on identifying such sections within patent documents from which compounds in SureChEMBL were captured.

The annotation pipeline within SureChEMBL leveraged numerous text and image mining tools for associating compounds to patents [26]. To enrich the provenance of the pipeline, it included the description of the section in the patent document from which the compound was captured. We leveraged this information to identify relevant sections in the patent document that mentioned the compound **(Figure 3)**. Our findings revealed that the chemical image is the primary source from which the compounds were extracted. This indicates that the patents linked to the compounds in SureChEMBL predominantly highlight the compound as the primary invention, whether related to its synthesis, formulation, or application within a specific area. Moreover, an equal amount of chemicals in the database were captured from the description section. In patent documents, the description section covers a vast range of text from the actual utility of the invention (in this case the compounds) to its methodology (for eg., formulation). Due to this vast context, there is the possibility to include compounds such as assay buffers and reactants which are important for the compound stability or assay protocol but not necessarily the main scope of the patent document. A small proportion of compounds are also annotated from the MOL file, which was one of the basic files that needed to be supplied with the chemical-related patents filed to the United States Patent and Trade Office (USPTO) (https://www.uspto.gov/web/offices/com/sol/og/2005/week28/patbiod.htm).

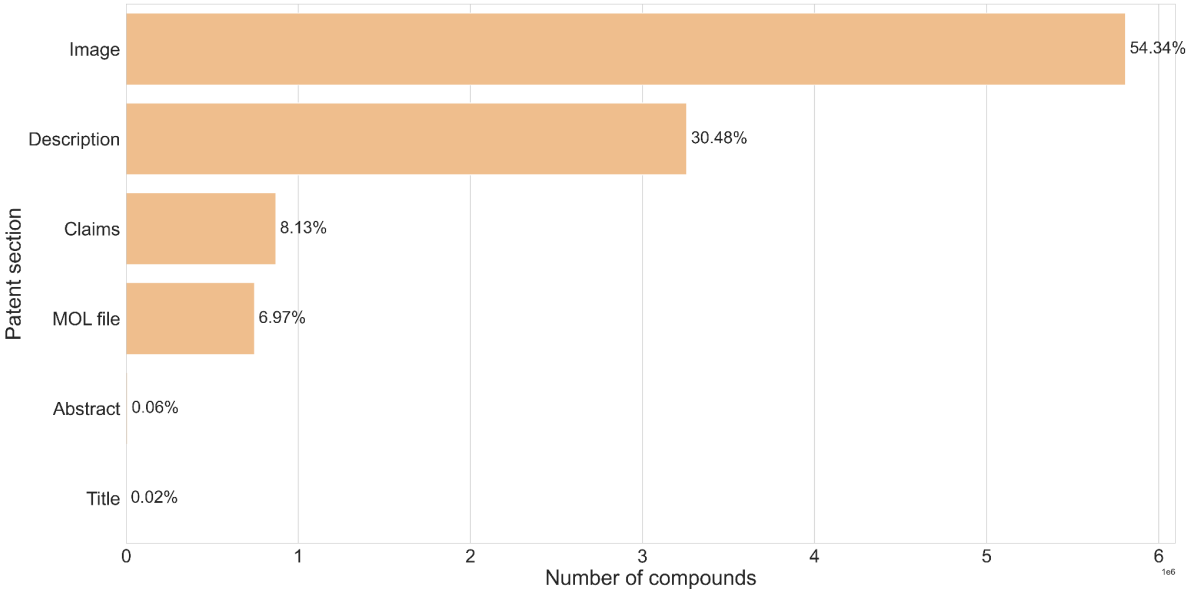


Figure 3: Prevalent sections in patent documents capturing compound information. The description section is the major source for compounds in SureChEMBL followed by textual data sources i.e. abstracts and claims sections. The data from image and MOL files were collected only for US patents after 2007.

3.3 Applicability of Ro5 and Beyond rules for chemical diversity in patents

In light of these two rules, Ro5 and bRo5, we investigated the classical characteristics of the compounds found within patent applications. In the past decade, a large contribution of Ro5 (75.94%) and bRo5 (24.06%) compounds have been found in patent documents. We looked into the top five most patented compounds that follow Ro5/bRo5 rules and found prominent compounds such as the antidepressant Vilazodone

(SCHEMBL650682) and generic ACE-inhibitor Captopril (SCHEMBL4) in the Ro5 category while Angiotensin-II analog Saralasin (SCHEMBL23) and natural products like Lankamycin (SCHEMBL1930864) in bRo5 category **(Figure 4)**.

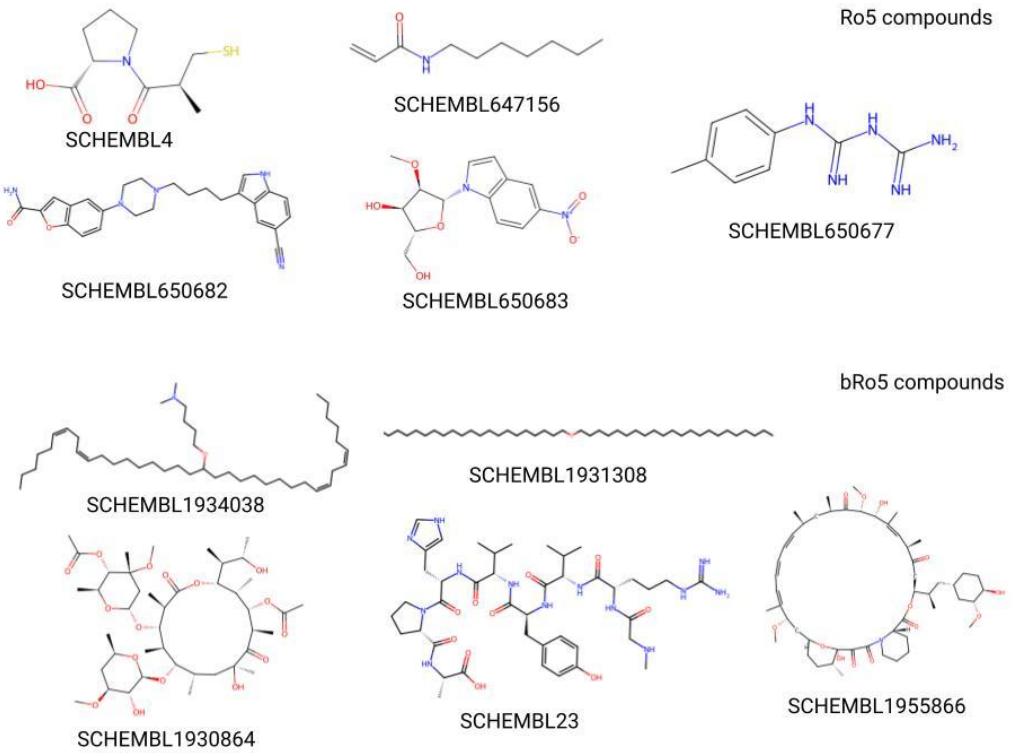


Figure 4: Top five compounds falling into the Ro5 and bRo5 categories. The top box covers Ro5 complaint compounds Captopril (SCHEMBL4), N-heptylacrylamide (SCHEMBL647156), 1-(p-tolyl)biguanide (SCHEMBL650677), 1-(2-O-Methyl-beta-D-ribofuranosyl)-5-nitro-1H-indole (SCHEMBL650683), Vilazodone (SCHEMBL650682) in clockwise manner. The bottom box covers the compounds beyond Ro5 space, including SCHEMBL1934038, Docosyl ether (SCHEMBL1931308), SCHEMBL1955866, Saralasin (SCHEMBL23), and Lankamycin (SCHEMBL1930864) in a clockwise manner.

Recently, PROteolysis-TArgeting Chimeras (PROTACS) have been identified to fall into the bRo5 category with the potential to progress into clinics [57-59]. PROTACs achieve protein degradation by "hijacking" the cell's ubiquitin-proteasome system (UPS) and bringing together the target protein and an E3 ligase. The UPS uses a complex process involving multiple steps. First, the E1 ligase activates and conjugates the ubiquitin to the E2 ligase; secondly, the E2 ligase forms a complex with the E3 ligase and the E3 ligase targets proteins and covalently attaches the ubiquitin to the protein of interest. Finally, after a ubiquitin chain is formed, the protein is recognized and degraded by the 26S proteasome. PROTACs take advantage of this cellular system by putting the protein of interest in close proximity to the E3 ligase, catalyzing its degradation. This mechanism has been successfully used in various applications, including cancer immunotherapy and infectious disease research [60]. Due to their non-adherent Ro5 characteristics, these molecules have not undergone “classical” prior optimization for oral bioavailability [61] and CNS penetration [62]. Additionally, in the past few years, interest has grown in the generation of macrocyclic compounds, those that retain the original scaffold or structure of existing compounds but contain additional function groups or side chains allowing for ring-shaped structures [63, 64]. With the increasing interest in such compounds as potential clinical and drug candidates, we looked into the calculated physicochemical properties of all patented compounds to check whether the chemical space expansion reflected PROTAC-like and macrocyclic compounds in recent years. **Table 2** shows the average physicochemical characteristics of compounds found annually. An increase in the average molecular weight, the hydrogen bond donor and acceptors, and the number of rings might be due to multiple factors including the interest in the development of PROTAC-like and/or macrocyclic molecules. Moreover, some common findings with respect to molecular properties across potential clinical candidates were found; for example, the consistent number of chiral centers (summarized in **Supplementary Table T1** and **Supplementary Figure S2)** mirrors

what was recently reported by Scott *et al*. (2022) [65]. Along the same lines, a growth in the number of hydrogen bond acceptors and rotatable bonds **(Supplementary Table T1)** and an increase in the number of rings could prove numerous factors alongside PROTACs, signaling macrocyclic compounds (e.g., cyclic peptides or cyclic kinase inhibitors) [66] or towards complex and diversifying synthesis capabilities [67, 68]. In a comprehensive view, these inclinations imply a pursuit of diverse structural characteristics, which we aimed to substantiate.

Year	Molecular weight	LogP	# HBA	# HBD	# NRotB	# Rings
2015	412.54	3.98	5.13	1.49	5.95	3.52
2016	442.26	4.43	5.42	1.55	6.33	3.90
2017	453.70	4.60	5.36	1.62	6.55	4.05
2018	454.57	4.67	5.38	1.53	6.54	4.07
2019	464.69	4.74	5.52	1.56	6.50	4.26
2020	475.60	4.97	5.59	1.54	6.54	4.47
2021	477.83	4.95	5.60	1.57	6.53	4.51
2022	498.71	5.43	5.71	1.50	6.71	4.94

Table 2: Molecular properties of compounds found in patents. For compounds found in each year, an average of the difference molecular properties: i) molecular weights, ii) LogP, iii) the number of hydrogen bond acceptors (# HBA), iv) the number of hydrogen bond donors (# HBD), v) the number of rotatable bonds (# NRotB), and vi) the number of any ring (# Rings) was taken.

Additionally, we were interested in the identification of compounds that could be covalent binders, showed existing polypharmacology, or were reactive in nature. One simple way to distinguish such compounds could be with the help of pan-assay interference compounds (PAINS) [69]. As a majority of the compounds were found with a MW < 2000 Da, we chose that as a cutoff to identify compounds that were flagged as PAINS **(Figure 5A)**. Of these compounds, about 3.7% compounds fall into 277 different PAINS alerts. The most prominent of these PAINS alerts include the presence of azo compounds (azo_A(324); 18.5%), the presence of compounds involving one aniline and two alkyl groups with either an additional alkyl group (anil_di_alk_A(478), 14%) or an addition carbon (anil_di_alk_C(246), 5.7%), the presence of compound containing an indole, a phenyl, and an alkyl group (indol_3yl_alk(461), 8.6%), and the presence of compounds containing catechol substructures (catechol_A(92), 6.1%). Aromatic PAINS like quinone also make it to the top of the list with about 5.4% compounds **(Supplementary Table S3)**. Additionally, compounds showing more than one PAINS alert were also found **(Figure 5B)**. The presence of these assay-interfering compounds is not surprising as a study by Capuzzi *et al.*, 2017 showed that FDA-approved drugs that contained PAINS were more active than the non-PAINS-containing drugs [70]. This has indeed led researchers like Senger *et al.* (2016) to question the filtering of promiscuous compounds during the early drug discovery steps [71]. Thus, acknowledging the presence of such PAINS liable compounds in patent documents is important when looking into the chemical space of these compounds. The question of whether these compounds are "bad" (showing false positive results in assays) or innocent (due to their possible polypharmacology activity) remains unsolved.

Lastly, we reduced the compounds to their Bemis-Murcko (BM) scaffold to quantify the diversity of compounds found in patents. The advantage of using a BM scaffold is two-fold: first, since these compounds are derived from patent documents, mapping them back to their original definition would provide clues on how they were derived and second, this representation framework retains the rings and side chains found in the compounds along with their atom types, unlike the graph framework that replaces all heteroatoms to carbon and collapses all bonds to single bonds notations [72]. In total, we found a total of 3 million distinct scaffolds encompassing our cohort of patent documents. Annually, this number reached about 0.4 million unique

scaffolds, which is equivalent to the amount of patents filed per year **(Figure 5C)**. Of these scaffolds, 43% were found to be associated with chemical images in patents while 42% were associated with patent description. Only 9.18% were found to be extracted from the patent claims sections **(Supplementary Table T4)**. A decline in the number of scaffolds generated in the past years has been observed but unexplored **(Figure 5C)**. This could be attributed to several reasons: some occasional factors like COVID dependent blockade of some patenting activities in 2022, or structural reasons rooted into medicinal chemistry syntheses which needs further and deeper analyses, like fading diversity of/in reagents or common and larger usage of X-ray target cavity information in the design of binders.

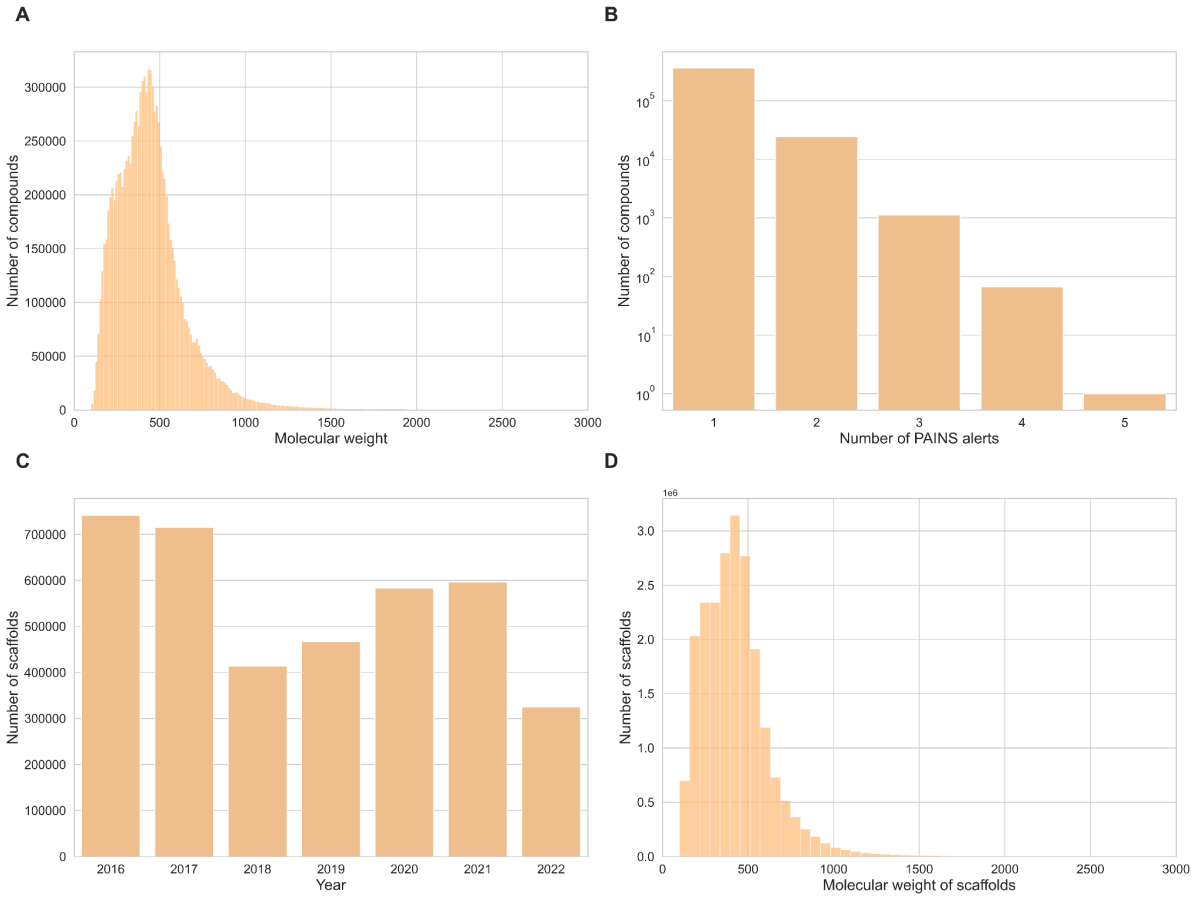


Figure 5: Summary of compounds based on Murcko scaffolds and PAINS alerts. A) Distribution of the molecular weight of the compounds. B) The number of PAINS alert structures found to be associated with compounds. C) The number of unique scaffolds found per year in patent documents from 2016 onwards. D) Molecular weight distribution of the scaffolds.

Reducing the compounds to their generic scaffolds would yield common compounds like benzene or furan. These common scaffolds might not exactly be found in patent documents but would be part of a larger scaffold shown in these documents. In our scaffold generation technique, we found a larger number of such common scaffolds when we looked into their respective molecular weight distribution **(Figure 5D)**. A larger number of these scaffolds have a single ring with variations in the side chains causing them to weigh about a few hundred daltons **(Figure 6)**. The most prominent ones include cyclic scaffolds such as cyclohexane and iso-propyl cyclohexane and linear scaffolds such as 2,3,8-trimethyl nonane. Additionally, we found good proportions of long-chain aliphatic compounds (more than 10 carbon atoms) like decanes (11%) annotated mainly from chemical images and from the description section of the patent document. The context surrounding the utility of such fatty acid chains would require patent lawyers and experts, which is out of the scope of the current manuscript.

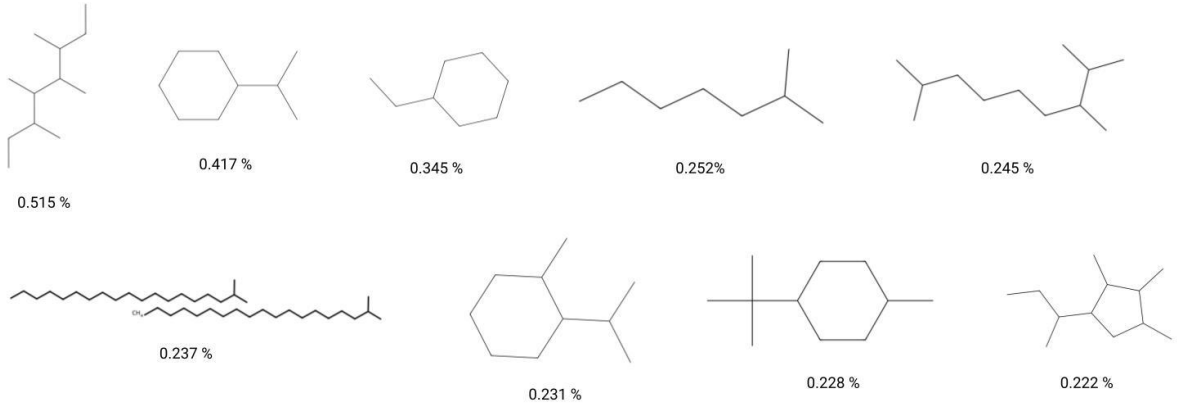


Figure 6: The top Murcko scaffold of compounds and their respective frequency found in patent documents. All these top scaffolds weigh less than 200 Daltons. The percentages only sum up to around 2.5% of the total, a clear sign of the structural diversity within the patent chemical space.

3.4 Traversing the trajectory of compounds from patent to clinical trials

Pre-established regulations like the Patent Act of 1990 or Trade-Related Aspects of Intellectual Property Rights (TRIPS) have aided drug proprietors in patenting novel pharmaceuticals or repurposing existing candidates to safeguard inventions under intellectual property laws before their integration into clinical practices [73, 74]. Consequently, many disparities have arisen between the drug names present in patent documents and their corresponding brand name (linked to their marketing), posing a challenge in finding associated patents [75]. In the past, successful endeavors were made to link drugs to patent documents by the FDA's Orange Book and the World Intellectual Property Organization (WIPO) through their Pat-INFORMED tool [76]. Acknowledging the complexity of the drug nomenclature across the different stages of drug development, we leveraged the chemical representation (InChIKey) of compounds in patents to generate an inventory of their corresponding clinical status.

DrugBank is one of the most comprehensive resources for chemical space exploration for compounds approved by the FDA and those under clinical trials. We started by looking into the overlap between the chemical space of compounds flagged by DrugBank as investigational and withdrawn with the compounds found in patents (i.e., SureChEMBL). By investigational drugs, DrugBank refers to compounds that are currently in clinical trials for a specific indication area (https://dev.drugbank.com/guides/terms/investigational-drug) while, withdrawn drugs refer to compounds that have been discontinued due to safety, toxicity, and/or adverse effects found in patients (https://dev.drugbank.com/guides/terms/withdrawn-drug). We found that a relatively small number of drugs (about 3,235) reached the clinics, with a mere 0.002% falling into the withdrawn drugs category as per DrugBank **(Figure 7A)**. Next, we delved into mapping the patented compounds with their research status (i.e., preclinical, clinical, and approved) according to what was reported in ChEMBL to identify the maximum clinical development stage the compounds in patents usually cover. **Figure 7B** depicts that compounds in patents are distributed across various research phases, ranging from preclinical to clinical, with the majority concentrated in the preclinical stage. Of the total compounds found in patents, roughly 1% of the drugs were approved for one or more indication areas. This is not surprising as the analysis by Brown (2023) showed that hit compounds evolve through the drug development and discovery pipeline as they undergo structural modification that ensures their reach into clinics [77]. Furthermore, a fraction of 1.6% of compounds had no information (classified as unknown by ChEMBL) regarding their clinical stage and were probably lost during translation from patents to clinics or failed to be captured and annotated by the resource database (in this case ChEMBL). Also, Phase 2 and Phase 3 exhibit a greater proportion of compounds than Phase 1, and this is attributed to the fact that these trial phases typically yield larger publications, assuming the trials were successful [78].

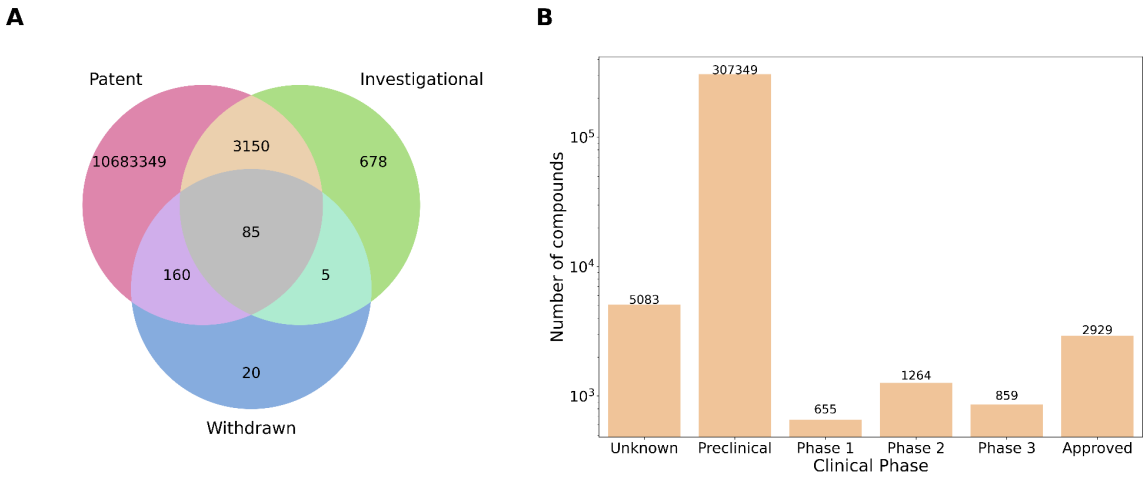


Figure 7: Overlap of the patented compounds on their drug status. A) Euler figure of the compounds found in patents against those classified as investigational and/or withdrawn drugs in DrugBank. B) Distribution of clinical phases of compounds found in ChEMBL.

3.5 Limitations

We acknowledge certain limitations in our analysis that deserve attention from the readers. Firstly, our analysis relies on open-source data, which may introduce potential data quality issues. For instance, periodic updates of data sources could lead to temporary gaps, possibly resulting in inaccuracies in our analysis, particularly in areas such as the clinical status of compounds discussed in our manuscript. Secondly, we assumed that all SureChEMBL patents are relevant to drug discovery, as they may pertain to disease indications or drug formulation. However, this assumption may not hold universally, and it would be preferable to refine our analysis by focusing on a subset of patent IPC codes, as outlined by Gadiya *et al*. (2023) [79], to create a more drug discovery-specific patent dataset. Lastly, we briefly look into drug repurposing patents and recognize the potential value in identifying specific indication areas where drug repurposing patents are concentrated. This could offer insights into the similarity of mode-of-action across different disease indications for certain compounds.

4. Discussion

Patent documents play an essential role in drug discovery and biological annotation pipelines such as those capturing the molecule image and converting them to SMILES and other chemical representations, or those highlighting gene and disease names in the patent documents that have emerged in the past few years. Here, we have focused on SureChEMBL, a publicly available patent database that leverages a number of proprietary and nonproprietary tools to collect, extract, and annotate patent documents in the life science domain, and inspected the data resource on the annotation quality for drug-like molecules and drug discovery related patent documents. Our results revealed that the chemical space represented in SureChEMBL can majorly be found in public chemical databases such as PubChem, a stronger collector of chemicals, compared to ChEMBL whose aim is to be a curated data resource with a focus on bioactivity dataset collection. Moreover, a very low proportion of chemical space was confined to SureChEMBL indicating the existence of compounds within chemical databases that are either held proprietary or have remained biologically undisclosed in ChEMBL by major pharmaceutical corporations. Furthermore, we also observed that the compounds found in SureChEMBL covered the Ro5 as well as bRo5 spaces hinting towards the cyclic molecules such as cyclic peptides, PROTACS, and PROTAC-like compounds in recent years. Following that, a detailed examination of the chemical activity associated with these compounds revealed the presence of promiscuous compounds such as PAINS within the database. This underscores the notion that the mere identification of compounds from patents falls short in the identification of drug leads or potential drug candidates; instead, it highlights the necessity for a medicinal chemistry-oriented filtration to eliminate non-active or activity-interfering compounds within such databases. Subsequently, we explored the Bemis-Murcko scaffolds of the compounds to analyze the chemical similarity

and diversity space. This exploration showed that the Murcko simplification is too generic to understand the diversity of the compounds as represented in patent documents, similar to Markush structures at least in the case of compounds annotated within patent documents. Furthermore, we found an equal number of Murcko scaffolds as patents per year indicating the increased diversity of compounds within patent documents. Finally, we also addressed the drug discovery path of a patent drug by looking into its transition from a patent to a clinic or market (i.e. post-approval). In conclusion, the manuscript provides a medicinal chemistry perspective on the chemical landscape formed by the patent compounds thus laying a foundation for future analysis.

Previously, discussion on the efficiency of patents for drug discovery has been questioned numerous times given their ability to not disclose the drug candidate, thereby protecting its IP. This manuscript enlightens the ability to transpire the data (specifically compounds) associated with patents by garnering data around those compounds. We found that the majority of the data found in these compounds are accessible through public resources, thus denoting the move towards transparent systems in the field of patents. This opens the future scope of the dataset for machine learning and artificially intelligent models in the future. For example, compounds found in patents could be used to enrich the veracity of datasets used for machine learning tasks by simply enriching the structure-activity relationship (SAR) landscape around the lead candidate. This can be a potential alternative to the generative AI-based approaches, provided that a patent around the lead molecule exists and has been previously mapped. Therefore, this huge resource has a vast potential that is yet to be mined and leverage for driving drug discovery purposes.

Author contribution

YG conceived the work. YG and AZ contributed to the ideation. YG and SS performed the analysis. MHA, PG, YG, and AZ have written the manuscript. All the authors have reviewed and approved the final manuscript.

Acknowledgments

We would like to thank the authors of the public resources used in our work for making their datasets available to the scientific community. The REMEDi4ALL project has received funding from the European Union's Horizon Europe Research & Innovation programme under grant agreement No 101057442. This work reflects only the authors' view and the EC is not responsible for any use that may be made of the information it contains.

Funding

This research was funded by EU Horizon Europe Research & Innovation programme REMEDi4ALL, grant no. 101057442.

References

1. Grabowski, H.G., DiMasi, J.A. and Long, G., 2015. The roles of patents and research and development incentives in biopharmaceutical innovation. *Health Affairs*, 34(2), pp.302-310. https://doi.org/10.1377/hlthaff.2014.1047
2. Kesselheim, A.S., Sinha, M.S. and Avorn, J., 2017. Determinants of market exclusivity for prescription drugs in the United States. *JAMA internal medicine*, 177(11), pp.1658-1664. https://doi.org/10.1001/jamainternmed.2017.4329
3. Dunn, Michael K. Timing of patent filing and market exclusivity. *Nature Reviews Drug Discovery* 10, no. 7 (2011): 487. https://doi.org/10.1038/nrd3494
4. Sayle, R.A., Petrov, P., Winter, J. and Muresan, S., 2011. Improved chemical text mining of patents using infinite dictionaries, translation and automatic spelling correction. *Journal of Cheminformatics*, 3, pp.1-1. https://doi.org/10.1186/1758-2946-3-S1-O16
5. Gadiya, Y., Gribbon, P., Hofmann-Apitius, M. et al., 2023. Pharmaceutical patent landscaping: A novel approach to understand patents from the drug discovery perspective. *Artificial Intelligence in the Life Sciences*, 3, p.100069. https://doi.org/10.1016/j.ailsci.2023.100069

6. Kong, X., Zuo, H., Huang, H.D., et al., 2023. STING as an emerging therapeutic target for drug discovery: Perspectives from the global patent landscape. *Journal of Advanced Research*, 44, pp.119-133. https://doi.org/10.1016/j.jare.2022.05.006
7. Zhang, H.L. and Li, Y., 2023. The Patent Landscape of BRAF Target and KRAS Target. *Recent Patents on Anti-Cancer Drug Discovery*, 18(4), pp.495-505. https://doi.org/10.2174/1574892818666221207091329
8. Song, C.H., Han, J.W., Jeong, B. and Yoon, J., 2017. Mapping the patent landscape in the field of personalized medicine. *Journal of Pharmaceutical Innovation*, 12, pp.238-248. https://doi.org/10.1007/s12247-017-9283-z
9. Lahiry, S. and Rangarajan, K., 2018. Patent Landscape for Indian Biopharmaceutical Sector: A Strategic Insight. *Flexible Strategies in VUCA Markets*, pp.31-47.
10. Mucke, Hermann. Intellectual Property Considerations. 20222022. https://doi.org/10.1039/9781839163401-00264
11. Strittmatter, S.M., 2014. Overcoming drug development bottlenecks with repurposing: old drugs learn new tricks. *Nature medicine*, 20(6), pp.590-591. https://doi.org/10.1038/nm.3595
12. Senger, S., 2017. Assessment of the significance of patent-derived information for the early identification of compound–target interaction hypotheses. *Journal of Cheminformatics*, 9(1), pp.1-8. https://doi.org/10.1186/s13321-017-0214-2
13. Colen, L., Belderbos, R., Kelchtermans, S. et al., 2023. Many are called, few are chosen: the role of science in drug development decisions. *The Journal of Technology Transfer*, pp.1-26. https://doi.org/10.1007/s10961-022-09982-6
14. Schmitt, V.J., Walter, L. and Schnittker, F.C., 2023. Assessment of patentability by means of semantic patent analysis–A mathematical-logical approach. *World Patent Information*, 73, p.102182. https://doi.org/10.1016/j.wpi.2023.102182
15. Fabry, B., Ernst, H., Langholz, J. et al., 2006. Patent portfolio analysis as a useful tool for identifying R&D and business opportunities—an empirical application in the nutrition and health industry. *World Patent Information*, 28(3), pp.215-225. https://doi.org/10.1016/j.wpi.2005.10.004
16. Grego, T., Pęzik, P., Couto, F.M. et al., 2009. Identification of chemical entities in patent documents. In *Distributed Computing, Artificial Intelligence, Bioinformatics, Soft Computing, and Ambient Assisted Living: 10th International Work-Conference on Artificial Neural Networks, IWANN 2009 Workshops*, Salamanca, Spain, June 10-12, 2009. Proceedings, Part II 10 (pp. 942-949). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-02481-8_144
17. Farre-Mensa, J., Hegde, D. and Ljungqvist, A., 2020. What is a patent worth? Evidence from the US patent "lottery". *The Journal of Finance*, 75(2), pp.639-682. https://doi.org/10.1111/jofi.12867
18. Falaguera, M.J. and Mestres, J., 2021. Identification of the core chemical structure in SureChEMBL patents. *Journal of Chemical Information and Modeling*, 61(5), pp.2241-2247. https://doi.org/10.1021/acs.jcim.1c00151
19. Kunimoto, R. and Bajorath, J., 2017. Exploring sets of molecules from patents and relationships to other active compounds in chemical space networks. *Journal of Computer-Aided Molecular Design*, 31, pp.779-788. https://doi.org/10.1007/s10822-017-0061-2
20. Wagner, S., Sternitzke, C. and Walter, S., 2022. Mapping Markush. *Research Policy*, 51(10), p.104597. https://doi.org/10.1016/j.respol.2022.104597
21. Deng, W., Berthel, S.J. and So, W.V., 2011. Intuitive patent Markush structure visualization tool for medicinal chemists. *Journal of chemical information and modeling*, 51(3), pp.511-520. https://doi.org/10.1021/ci100261u
22. Wills, T.J. and Lipkus, A.H., 2020. Structural approach to assessing the innovativeness of new drugs finds accelerating rate of innovation. *ACS Medicinal Chemistry Letters*, 11(11), pp.2114-2119. https://doi.org/10.1021/acsmedchemlett.0c00319
23. Kim, J. and Lee, S., 2015. Patent databases for innovation studies: A comparative analysis of USPTO, EPO, JPO and KIPO. *Technological Forecasting and Social Change*, 92, pp.332-345. https://doi.org/10.1016/j.techfore.2015.01.009

24. Marco, A.C., Myers, A., Graham, S.J., et al., 2015. The USPTO patent assignment dataset: Descriptions and analysis. http://dx.doi.org/10.2139/ssrn.2849634
25. Hill, L.L., 2005. The orange book. *Nature Reviews*. Drug Discovery, 4(8), p.621. https://doi.org/10.1038/nrd1804
26. Papadatos, George, Mark Davies, Nathan Dedman, et al. SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic acids research* 44, no. D1 (2016): D1220-D1228. https://doi.org/10.1093/nar/gkv1253
27. Kim, Sunghwan, Jie Chen, Tiejun Cheng, et al. PubChem 2023 update. *Nucleic Acids Research* 51, no. D1 (2023): D1373-D1380. https://doi.org/10.1093/nar/gkac956
28. Gaulton, Anna, Louisa J. Bellis, A. Patricia Bento, et al. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* 40, no. D1 (2012): D1100-D1107. https://doi.org/10.1093/nar/gkr777
29. Wishart, David S., Yannick D. Feunang, An C. Guo, et al. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic acids research* 46, no. D1 (2018): D1074-D1082. https://doi.org/10.1093/nar/gkx1037
30. Pollastri, Michael P. Overview on the Rule of Five. *Current protocols in pharmacology* 49, no. 1 (2010): 9-12. https://doi.org/10.1002/0471141755.ph0912s49
31. Lipinski, Christopher A., Franco Lombardo, Beryl W. Dominy, et al. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews* 64 (2012): 4-17. https://doi.org/10.1016/j.addr.2012.09.019
32. Lipinski, Christopher A. Drug-like properties and the causes of poor solubility and poor permeability. *Journal of pharmacological and toxicological methods* 44, no. 1 (2000): 235-249. https://doi.org/10.1016/S1056-8719(00)00107-6
33. Veber, Daniel F., Stephen R. Johnson, Hung-Yuan Cheng, et al. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of medicinal chemistry* 45, no. 12 (2002): 2615-2623. https://doi.org/10.1021/jm020017n
34. Doak, Bradley Croy, Björn Over, Fabrizio Giordanetto, et al. Oral druggable space beyond the rule of 5: insights from drugs and clinical candidates. *Chemistry & biology* 21, no. 9 (2014): 1115-1142. https://doi.org/10.1016/j.chembiol.2014.08.013
35. Terrett, N., 2013. Drugs in middle space. *MedChemComm*, 4(3), pp.474-475. https://doi/org/10.1039/C2MD90062A
36. Abad-Zapatero, Cele. A Sorcerer's apprentice and the rule of five: from rule-of-thumb to commandment and beyond. *Drug Discovery Today* 12, no. 23 (2007): 995-997. https://doi.org/10.1016/j.drudis.2007.10.022
37. Doak, Bradley C., Jie Zheng, Doreen Dobritzsch, et al. How beyond rule of 5 drugs and clinical candidates bind to their targets. *Journal of medicinal chemistry* 59, no. 6 (2016): 2312-2327. https://doi.org/10.1021/acs.jmedchem.5b01286
38. Bemis, G.W. and Murcko, M.A., 1996. The properties of known drugs. 1. Molecular frameworks. *Journal of medicinal chemistry*, 39(15), pp.2887-2893. https://doi.org/10.1021/jm9602928
39. Baell, J. and Walters, M.A., 2014. Chemistry: Chemical con artists foil drug discovery. *Nature*, 513(7519), pp.481-483. https://doi.org/10.1038/513481a
40. Landrum, Greg. RDKit: A software suite for cheminformatics, computational chemistry, and predictive modeling. Greg Landrum 8 (2013). https://github.com/rdkit/rdkit
41. Swain, Matt. PubChemPy: A way to interact with PubChem in Python. (2014). https://pubchempy.readthedocs.io/
42. Senger, Stefan, Luca Bartek, George Papadatos, and Anna Gaulton. Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents. *Journal of cheminformatics* 7, no. 1 (2015): 1-12. https://doi.org/10.1186/s13321-015-0097-z
43. EPO. Numbering system - Japan. *EPO*. Last modified May 7, 2021. Last accessed August 22, 2023 https://www.epo.org/searching-for-patents/helpful-resources/asian/japan/numbering.html.
44. Ferrence, Gregory M., Clare A. Tovee, Stephen JW Holgate, et al. CSD Communications of the Cambridge Structural Database. *IUCrJ* 10, no. 1 (2023). https://doi.org/10.1107/S2052252522010545

45. Southan, C., Sitzmann, M. and Muresan, S. (2013), Comparing the Chemical Structure and Protein Content of ChEMBL, DrugBank, Human Metabolome Database and the Therapeutic Target Database. *Mol. Inf.*, 32: 881-897. https://doi.org/10.1002/minf.201300103
46. Sauban Ghani, S. (2020). A comprehensive review of database resources in chemistry. *Eclética Química*, 45(3), 57–68. https://doi.org/10.26850/1678-4618eqj.v45.3.2020.p57-68
47. Tamura, Shunsuke, Tomoyuki Miyao, and Jürgen Bajorath. Large-scale prediction of activity cliffs using machine and deep learning methods of increasing complexity. *Journal of Cheminformatics* 15, no. 1 (2023): 1-11. https://doi.org/10.1186/s13321-022-00676-7
48. Tran, Thi Tuyet Van, Agung Surya Wibowo, Hilal Tayara, and Kil To Chong. Artificial Intelligence in Drug Toxicity Prediction: Recent Advances, Challenges, and Future Perspectives. *Journal of Chemical Information and Modeling* (2023). https://doi.org/10.1021/acs.jcim.3c00200
49. Lagunin, Alexey A., Anastasia V. Rudik, Pavel V. Pogodin, et al. CLC-Pred 2.0: A Freely Available Web Application for In Silico Prediction of Human Cell Line Cytotoxicity and Molecular Mechanisms of Action for Druglike Compounds. *International Journal of Molecular Sciences* 24, no. 2 (2023): 1689. https://doi.org/10.3390/ijms24021689
50. Chen, Wei, Xuesong Liu, Sanyin Zhang, and Shilin Chen. Artificial intelligence for drug discovery: Resources, methods, and applications. *Molecular Therapy-Nucleic Acids* (2023). https://doi.org/10.1016/j.omtn.2023.02.019
51. Bhattacharjee, Apurba K. Pharmacophore-based virtual screening of large compound databases can aid "big data" problems in drug discovery. *In Big Data Analytics in Chemoinformatics and Bioinformatics*, pp. 231-246. Elsevier, 2023. https://doi.org/10.1016/B978-0-323-85713-0.00014-1
52. Almansour, Nahlah Makki, Khaled S. Allemailem, Abeer Abas Abd El Aty, et al. In Silico Mining of Natural Products Atlas (NPAtlas) Database for Identifying Effective Bcl-2 Inhibitors: Molecular Docking, Molecular Dynamics, and Pharmacokinetics Characteristics. *Molecules* 28, no. 2 (2023): 783. https://doi.org/10.3390/molecules28020783
53. Ohms, Joerg. Current methodologies for chemical compound searching in patents: A case study. *World Patent Information* 66 (2021): 102055. https://doi.org/10.1016/j.wpi.2021.102055
54. Falaguera, Maria J., and Jordi Mestres. Congenericity of claimed compounds in patent applications. *Molecules* 26, no. 17 (2021): 5253. https://doi.org/10.3390/molecules26175253
55. Structure and Function of the patent specification. HGF. https://www.hgf.com/wp-content/uploads/2020/09/Structure-and-Function-of-the-Patent-Specification.pdf (16 June 2023, last accessed)
56. Jessop, David M., Sam E. Adams, and Peter Murray-Rust. Mining chemical information from open patents. *Journal of cheminformatics* 3, no. 1 (2011): 1-17. https://doi.org/10.1186/1758-2946-3-40
57. Ermondi, Giuseppe, Diego Garcia Jimenez, Matteo Rossi Sebastiano, et al. Rational control of molecular properties is mandatory to exploit the potential of PROTACs as oral drugs. *ACS Medicinal Chemistry Letters* 12, no. 7 (2021): 1056-1060. https://doi.org/10.1021/acsmedchemlett.1c00298
58. Edmondson, Scott D., Bin Yang, and Charlene Fallan. Proteolysis targeting chimeras (PROTACs) in 'beyond rule-of-five'chemical space: Recent progress and future challenges. *Bioorganic & Medicinal Chemistry Letters* 29, no. 13 (2019): 1555-1564. https://doi.org/10.1016/j.bmcl.2019.04.030
59. García Jiménez, Diego, Matteo Rossi Sebastiano, Maura Vallaro, et al. Designing soluble PROTACs: strategies and preliminary guidelines. *Journal of Medicinal Chemistry* 65, no. 19 (2022): 12639-12649. https://doi.org/10.1021/acs.jmedchem.2c00201
60. Nalawansha, Dhanusha A., and Craig M. Crews. PROTACs: an emerging therapeutic modality in precision medicine. *Cell chemical biology* 27, no. 8 (2020): 998-1014. https://doi.org/10.1016/j.chembiol.2020.07.020
61. Ermondi, Giuseppe, Diego Garcia-Jimenez, and Giulia Caron. Protacs and building blocks: The 2d chemical space in very early drug discovery. *Molecules* 26, no. 3 (2021): 672. https://doi.org/10.3390/molecules26030672
62. Tashima, T., 2023. Proteolysis-Targeting Chimera (PROTAC) Delivery into the Brain across the Blood-Brain Barrier. *Antibodies*, 12(3), p.43. https://doi.org/10.3390/antib12030043

63. Xie, J. and Bogliotti, N., 2014. Synthesis and applications of carbohydrate-derived macrocyclic compounds. *Chemical reviews*, 114(15), pp.7678-7739. https://doi.org/10.1021/cr400035j
64. Zhao, Z. and Bourne, P., 2023. Rigid Scaffolds Are Promising for Designing Macrocyclic Kinase Inhibitors. *ACS Pharmacology & Translational Science*, 6 (8), pp. 1182-1191. https://doi.org/10.1021/acsptsci.3c00078
65. Scott, Kevin A., Nathalie Ropek, Bruno Melillo, et al. Stereochemical diversity as a source of discovery in chemical biology. *Current Research in Chemical Biology* 2 (2022): 100028. https://doi.org/10.1016/j.crchbi.2022.100028
66. Guo, Yukuang, Shaofei Ji, Wei Wang, et al. An integrated strategy for assessing the metabolic stability and biotransformation of macrocyclic peptides in drug discovery toward oral delivery. *Analytical Chemistry* 94, no. 4 (2022): 2032-2041. https://doi.org/10.1021/acs.analchem.1c04008
67. Münzfeld, Luca., Gillhuber, Sebastian., Hauser, Adrian. et al. Synthesis and properties of cyclic sandwich compounds. *Nature* 620, 92–96 (2023). https://doi.org/10.1038/s41586-023-06192-4
68. Xing, Gao., Miaoren, Xia., Chunhao, Yuan,. et al. Enantioselective Synthesis of Chiral Medium-Sized Cyclic Compounds via Tandem Cycloaddition/Cope Rearrangement Strategy. *ACS Catalysis* 2019 9 (3), 1645-1654. https://doi.org/10.1021/acscatal.8b04590
69. Chakravorty, S.J., Chan, J., Greenwood, M.N., et al., 2018. Nuisance compounds, PAINS filters, and dark chemical matter in the GSK HTS collection. *SLAS Discovery: Advancing Life Sciences R&D*, 23(6), pp.532-545. https://doi.org/10.1177/2472555218768497
70. Capuzzi, S.J., Muratov, E.N. and Tropsha, A., 2017. Phantom PAINS: Problems with the Utility of Alerts for Pan-Assay INterference Compound S. *Journal of chemical information and modeling*, 57(3), pp.417-427. https://doi.org/10.1021/acs.jcim.6b00465
71. Senger, M.R., Fraga, C.A., Dantas, R.F. et a.l, 2016. Filtering promiscuous compounds in early drug discovery: is it a good idea?. *Drug Discovery Today*, 21(6), pp.868-872. https://doi.org/10.1016/j.drudis.2016.02.004
72. Langdon, S.R., Brown, N. and Blagg, J., 2011. Scaffold diversity of exemplified medicinal chemistry space. *Journal of chemical information and modeling*, 51(9), pp.2174-2185. https://doi.org/10.1021/ci2001428
73. World Trade Organization (WTO), Agreement on Trade-Related Aspects of Intellectual Property Rights, Apr. 15, 1994, Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, 1869 U.N.T.S. 299, 33 I.L.M. 1197 (1994).
74. Motari, M., Nikiema, J.B., Kasilo, O.M., et al., 2021. The role of intellectual property rights on access to medicines in the WHO African region: 25 years after the TRIPS agreement. *BMC Public Health*, 21(1), pp.1-19. https://doi.org/10.1186/s12889-021-10374-y
75. Thakkar, K.B. and Billa, G., 2013. The concept of: Generic drugs and patented drugs vs. brand name drugs and non-proprietary (generic) name drugs. *Frontiers in Pharmacology*, 4, p.113. https://doi.org/10.3389/fphar.2013.00113
76. Schultz, M., 2018. Pat-INFORMED: a new tool for drug procurement. *WIPO MAGAZINE*, (5), pp.30-36. https://search.informit.org/doi/10.3316/agispt.20181002002556
77. Brown, D.G., 2023. An Analysis of Successful Hit-to-Clinical Candidate Pairs. *Journal of Medicinal Chemistry.* https://doi.org/10.1021/acs.jmedchem.3c00521
78. Cuschieri, Sarah. Clinical trial publications. *Saudi Journal of Anaesthesia* 13(Suppl 1), p S42-S43. https://doi.org/10.4103/sja.SJA_575_18
79. Gadiya, Y., Zaliani, A., Gribbon, P. and Hofmann-Apitius, M., 2023. PEMT: a patent enrichment tool for drug discovery. *Bioinformatics*, 39(1), p.btac716. https://doi.org/10.1093/bioinformatics/btac716

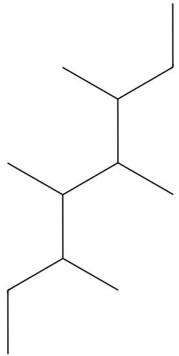
Author Description (30 words)

- Yojana Gadiya is a Ph.D. researcher and data scientist specializing in working with knowledge graphs and their applications in drug discovery along with patent literature mining.

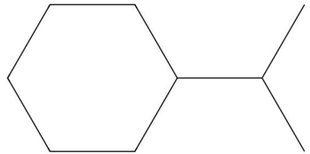
- Simran Shetty was a student intern at Fraunhofer ITMP during this work and is currently pursuing her master's degree in pharmaceutical sciences.
- Dr. Martin Hofmann-Apitutus is the Head of the Department of Bioinformatics at Fraunhofer SCAI. His scientific work focuses on information extraction, knowledge graphs, and modeling neurodegenerative and neuropsychiatric diseases.
- Dr. Philip Gribbon is the head of the department at Fraunhofer ITMP with expertise in pre-clinical screening data generation and analysis.
- Andrea Zaliani is a senior data scientist with an experience of more than three decades as a medicinal chemist. His current interests lie in chemoinformatics and data science.

Key Points

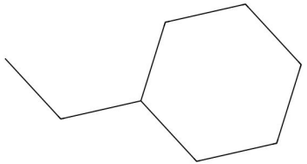
- In recent years, there has been an increasing interest in understanding and extracting information from patent documents for pharmaceutical drug discovery.
- We evaluate the SureChEMBL database from a research and development (R&D) and medicinal chemistry perspective to assess the characteristics of the chemical space found in patents for drug discovery.
- We demonstrate that compounds from patent documents can assist drug developers in accessing new chemical spaces and help researchers filter their drug candidates for further development.



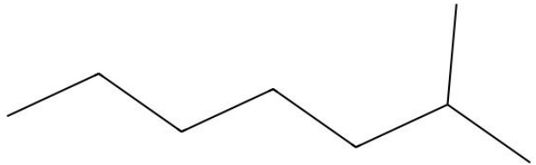
0.515 %



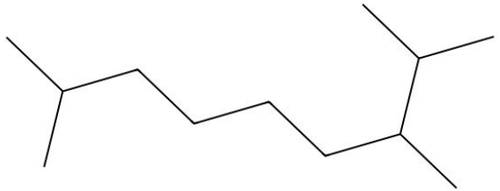
0.417 %



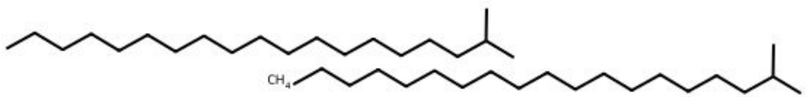
0.345 %



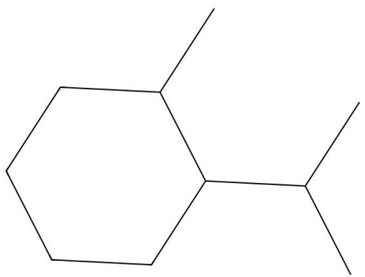
0.252%



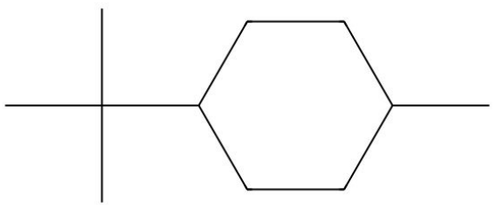
0.245 %



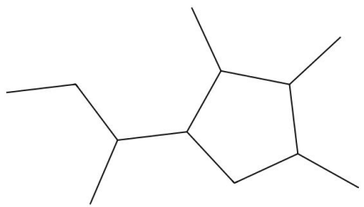
0.237 %



0.231 %

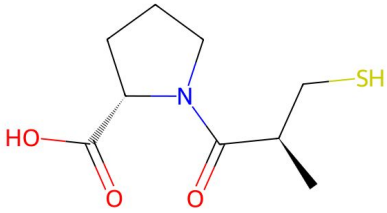


0.228 %

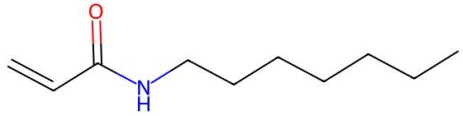


0.222 %

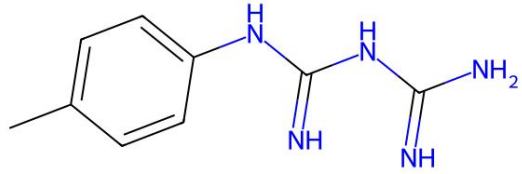
Ro5 compounds



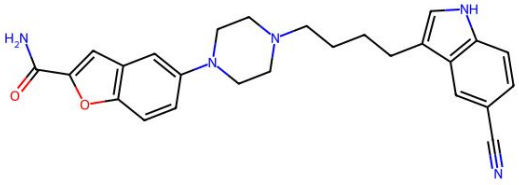
SCHEMBL4



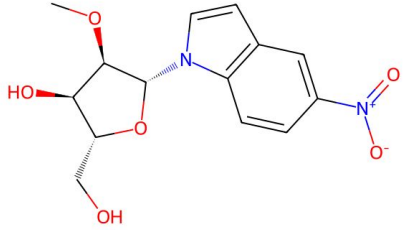
SCHEMBL647156



SCHEMBL650677

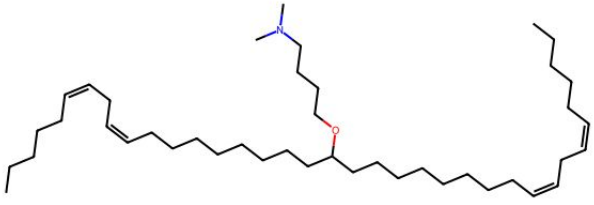


SCHEMBL650682



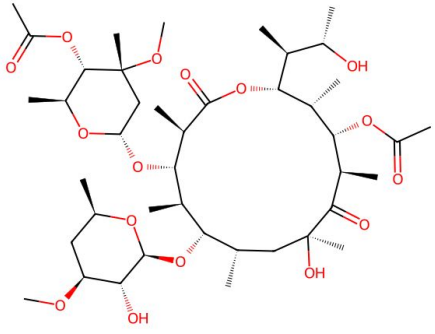
SCHEMBL650683

bRo5 compounds

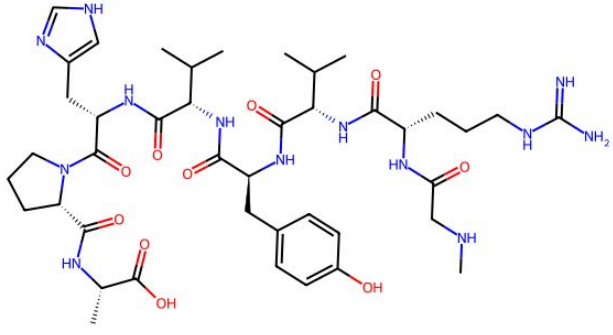


SCHEMBL1934038

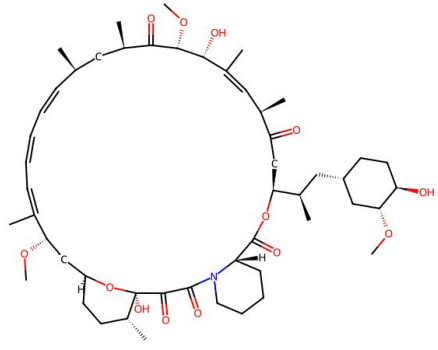
SCHEMBL1931308



SCHEMBL1930864



SCHEMBL23



SCHEMBL1955866