scientific data

OPEN

ANALYSIS

Exploring SureChEMBL from a drug discovery perspective

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

In the pharmaceutical industry, the patent protection of drugs and medicines is accorded importance because of the high costs involved in the development of novel drugs. Over the years, researchers have analyzed patent documents to identify freedom-to-operate spaces for novel drug candidates. To assist this, several well-established public patent document data repositories have enabled automated methodologies for extracting information on therapeutic agents. In this study, we delve into one such publicly available patent database, SureChEMBL, which catalogues patent documents related to life sciences. Our exploration begins by identifying patent compounds across public chemical data resources, followed by pinpointing sections in patent documents where the chemical annotations were found. Next, we exhibit the potential of compounds to serve as drug candidates by evaluating their conformity to drug-likeness criteria. Lastly, we examine the drug development stage reported for these compounds to understand their clinical success. In summary, our investigation aims at providing a comprehensive overview of the patent compounds catalogued in SureChEMBL, assessing their relevance to pharmaceutical drug discovery.

Introduction

Patent documents 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 document generally has the exclusive right to make, use, and sell the invention for approximately 20 years in a given jurisdiction[1,2]. In drug discovery, researchers explore patent documents to identify competing interests associated with a drug candidate across various organizations, such as pharmaceutical companies, universities, or individuals[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 therapeutic design, synthesis, and formulation, among others claims, 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 patent documents 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/generic name, etc. Patent documents 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 and professionals 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 document is usually found in the claims or description sections of the document 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 document, making it a subject of study for patent lawyers and pharma R&D scientists who seek to comprehend the FTO space associated with the patent documents.

[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. ✉e-mail: Yojana.Gadiya@itmp.fraunhofer.de

Pharmaceutical patent analysis covers a wide range of topics, including patenting trends[13], tools for patent protection[14,15] and the identification of novel chemical entities for certain treatments or diseases[16,17]. In a study by Falaguera and Mestres[18,19], compounds mined from SureChEMBL, a public patent database, were found to be a collection of starting materials, intermediate products, or pharmacologically relevant compounds (i.e. compounds that target genes or diseases)[18]. In order to make these compound classifications, the authors employed chemoinformatic methods, such as the matched molecular pair (MMP) analysis[20] and the maximum common substructure (MCS) search[18], as alternatives to the generic Markush structural searches[21,22]. Additionally, these methods have allowed for the generation of metrics to assess the chemical novelty and patentability of new compounds[23]. Despite 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 can be attributed to three main reasons: (i) the United States is the world's largest pharmaceutical market[24], (ii) the presence of an easy-to-use US-centric public patent database, the United States Patent and Trademark Office (USPTO), which allows for bulk download of patent documents and their metadata[25], and (iii) the availability of resources, such as the FDA's Orange Book, which allows for the tracking of drug candidates and their corresponding patent documents through time[26]. Furthermore, these analyses restrict pharmaceutical patent documents to those tagged with International Patent Classification (IPC) code A61K, an IPC class that includes hygiene-related patent documents in addition to medicinal ones, potentially merging non-pharmaceutical annotations.

SureChEMBL (https://www.surechembl.org/) is an extensive publicly available patent compound data catalogue for the life sciences[27]. This database identifies compounds, along with other biomedical entities, such as genes and diseases, from patent documents through the use of automated text and image mining pipelines. Furthermore, SureChEMBL keeps an individual record of each extracted compound, associating it with structural information (i.e., SMILES and InChIKeys) and the section of the patent document (i.e. claims, title, description, etc.) where the compound was extracted from. In this study, we aim to investigate the relevance of compounds annotated by SureChEMBL's pipeline with respect to approved drugs in the pharmaceutical market. Specifically, we applied a medicinal chemistry lens on compounds in SureChEMBL to identify patterns within their molecular scaffolds, as well as the physiochemical properties of patented compounds. Moreover, we assessed the similarity of these compounds to drugs through drug-likeness traits defined by Lipinski (Rule of Five). Rather than limiting the investigation to patent documents found in the United States, as done in all previous methodologies, we broaden our scope to a diverse dataset of patent applications filed and/or granted globally. By doing so, we covered larger IPC patent classes, including information on the medicinal utility of compounds and their formulations. Furthermore, our exploration scrutinizes the availability of compounds described in patent documents and beyond, specifically those annotated by public compound databases. We conclude by delving into the clinical candidate space of the patent compounds in order to understand the success rate of progressing compounds from patent application to clinical practice.

Results

In the following subsections, we evaluate the chemical space of patent documents found in SureChEMBL, an open-access public patent database. First, we provide a brief statistical summary of the data present in SureChEMBL, with a focus on the country where the patent documents were first registered. Afterwards, we discuss the searchability (i.e. the ability to search for compounds in other databases) of the patent compounds within large chemical databases, namely PubChem, ChEMBL and DrugBank. Next, we review the findability (i.e. the ability to identify the section in patent documents through which the compounds were annotated) of the patent compounds. Following this, we explored the drug-likeness of patent compounds through rules like Ro5 and beyond, evaluated their structural diversity through the Murcko scaffold, and reviewed the presence of structural alerts like Pan-assay interference structures (PAINS) in these compounds. Finally, we briefly discuss the progression of a compound from patent documents to the market through clinical trials.

Quantitative overview of data in SureChEMBL. From the statistical side, our dataset included a collection of 10 million compounds found in over 1.5 million patent applications (including both granted and non-granted) between 2015 and 2022. Throughout this study, we used the term "patent compounds" to identify those compounds that were captured and annotated by SureChEMBL's internal pipeline to be associated with a patent application. The patent documents in SureChEMBL are captured across a number of patent offices, namely the USPTO, European Patent Office (EPO), Japan Patent Office (JPO) and World Intellectual Property Organization (WIPO). However, it is worth noting that WIPO usually consists of only filed patent documents and does not have the authority to grant any patent. Additionally, the patent documents in SureChEMBL cover a broad range of IPC classes, such as human necessities (A01, A23, A24, A61, A62B), chemistry and metallurgy-oriented (C05, C06, C07, C08, C09, C10, C11, C12, C13, C14) and physics (G01N), all of which are part of this study.

In SureChEMBL, a patent document is assigned a unique SureChEMBL patent number (SCPN) (for eg. US-1234567-A1) that consists of a country code (based on the country the patent document was first registered in), followed by a 7–11 digit number, and a patent kind code. In our study, we used the country code in the SCPN to understand the distribution of patent documents across different patent offices. This exploration revealed that patent documents were predominantly filed in the United States and Europe, with 57.3% (24% granted) and 26.6% (11% granted) patents, respectively. Moreover, SureChEMBL also aggregated compounds from Japan (JP), but we found the contribution of these patent documents to be less than 1%[27] (Fig. 1a).

As mentioned previously, each patent document is associated with a "patent kind code" by SureChEMBL. This code is a two-letter alphanumeric code that assists patent officers and reviewers in efficiently distinguishing different kinds of patent documents, such as utility, design, or plant patents. Utility patent documents involve the "discovery and invention of new and useful processes", design patent documents cover the invention of a "novel design for an article of manufacture", and plant patent documents cover the scope for "discovering an asexually

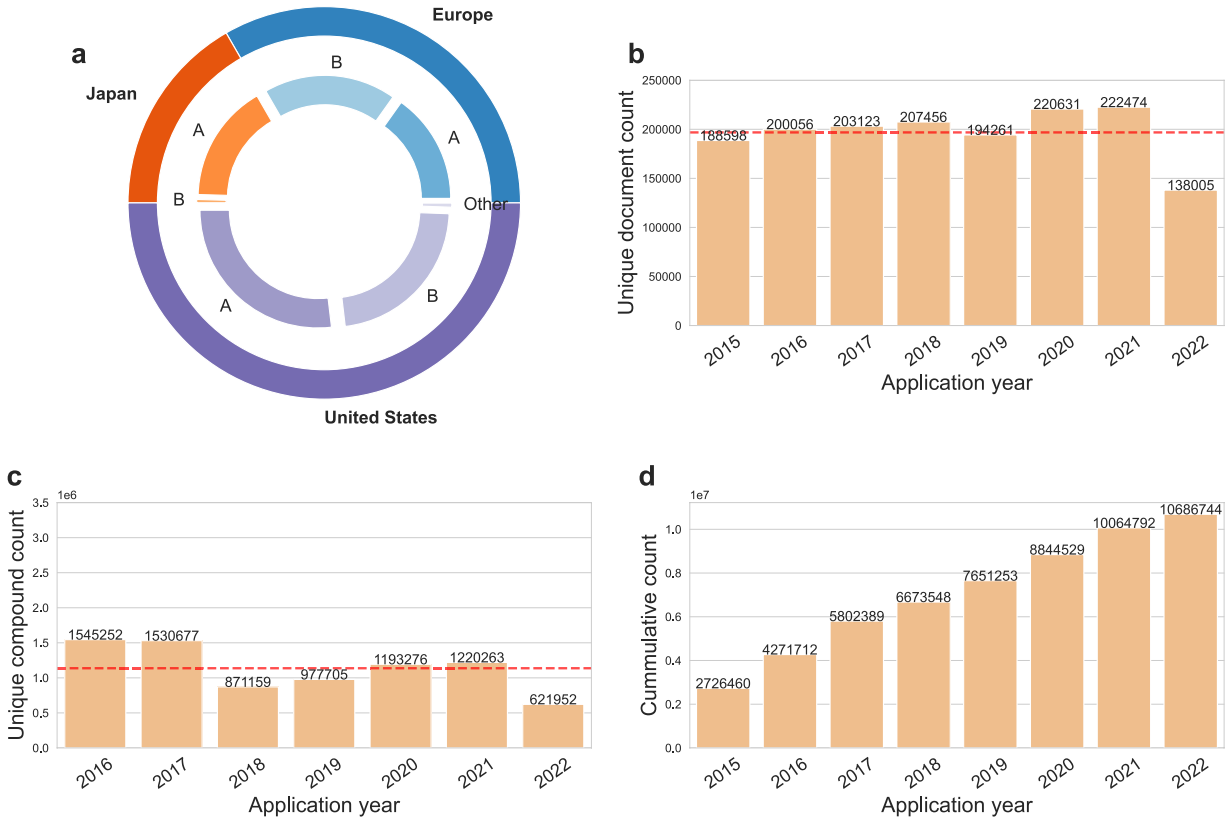

Fig. 1 (**a**) Distribution of patent document types by their patent kind codes across countries. (**b**) Distribution of patent documents filed over the application years. (**c**) Distribution of patent compounds over the application years. (**d**) Cumulative compound count over the years. The red line in subplots b and c indicates the average number of patents and compounds, respectively. Additionally, the counts displayed in subplots b and c are deduplicated counts for patents documents and compounds respectively.

Country	Patent kind code	Number of patent compounds
United States (US)	Filed (A)	6,111,699
United States (US)	Granted (B)	5,207,524
United States (US)	Design Patent (S)	2
United States (US)	Reissue Patent (E)	46,215
United States (US)	Plant Patent (P)	402
Europe (EP)	Filed (A)	2,648,617
Europe (EP)	Granted (B)	3,173,776
Japan (JP)	Filed (A)	254
Japan (JP)	Granted (B)	7

Table 1. Summary of the number of compounds found with respect to patent kind and country of filing. The compounds were counted based on their unique InChIKey representation in SureChEMBL.

reproducing variety of plant". To understand the proportion of these three different patent document types in SureChEMBL, we studied the patent kind codes for each of the registered patent documents in each jurisdiction individually. We identified large proportions of utility patent documents, including both filed (indicated by kind code A*X*) and granted patents (indicated by kind code B*X*) in each of the three countries (i.e. the United States, Europe and Japan), with the prior patent document type being predominant (Fig. 1a). Additionally, in the United States, we found the presence of a small proportion of "other" patent document classes, such as design patent documents (indicated by kind code S*X*), reissued patent documents (indicated by kind code E*X*) and plant patent documents (indicated by kind code P*X*) (summarized in Table 1).

Next, we investigated the distribution of patent documents and their compounds yearly. To do so, we collected all the patent documents and the patent compounds in SureChEMBL and distinguished them based on their application number and InChIKeys, respectively. On average, we found that 196,826 patent applications have been filed and patented each year (Fig. 1b,c). Moreover, an average of 6 compound references were

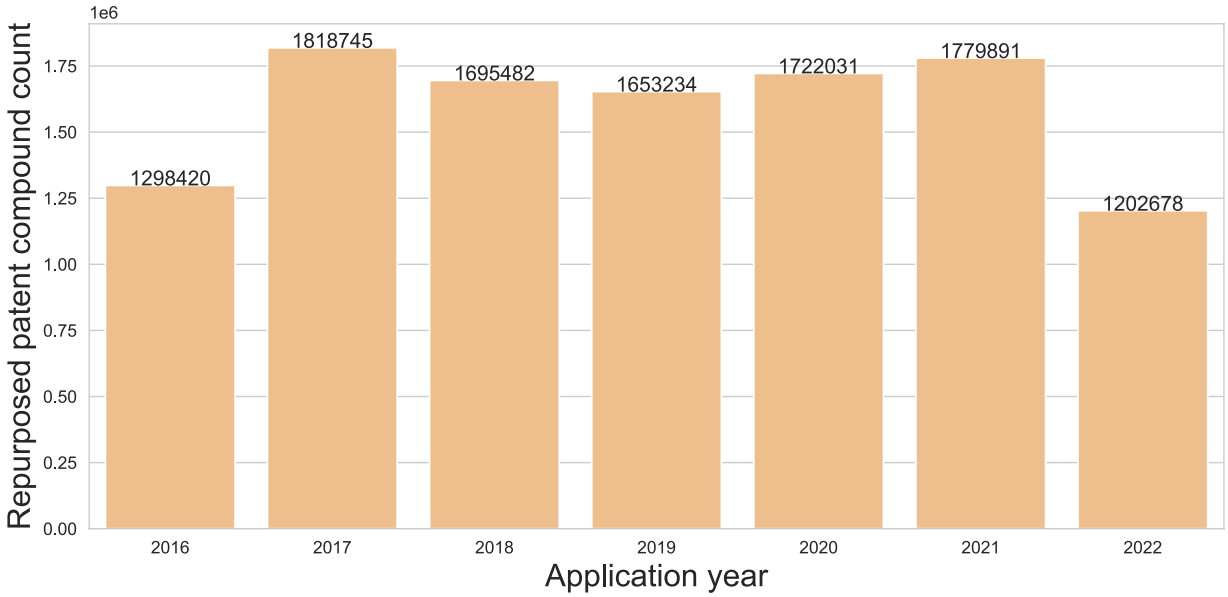

Fig. 2 Repurposed patent compound distribution over the application years. Distribution of patent compounds that have been found in patent documents from previous application years. It is to be noted that the repurposing scenario shown here is only considered from 2015 onwards.

identified per patent document in SureChEMBL. In addition to this, we examined the occurrence of patent compounds in more than one patent document. This analysis illustrated that, of the nearly 1 million patent compounds, 0.2% were associated with multiple patent documents (Fig. 2). A detailed investigation revealed that the majority (95%; 10,148,500 of 10,686,744) of these patent compounds appeared in fewer than 5 patent documents. On the other hand, 11,613 (of 10,686,744) patent compounds were found to be promiscuous across patent documents with their appearance in more than 1,000 documents each.

PubChem demonstrates highest coverage of patent compounds. A large number of compound-centric biological databases have been established in the past decades[28–30]. These databases have served various purposes in drug discovery, from identifying the bioactivity of unknown compounds[31,32] to the prediction of mechanisms of action[33,34], or simply for the virtual screening of drug candidates[35,36]. While previous research by Joerg Ohms (2022) explored the coverage of patent documents in relation to chemicals in two patent databases (SureChEMBL and Patentscope), the scope of this study was limited to manually comparing a set of chemicals between PubChem Substance and patent compounds[37]. Thus, to systematically understand the coverage of patent compounds in prominently used chemical databases, we analyzed the structural overlap between the compounds cited in patent documents and those found in three public chemical databases, namely PubChem, ChEMBL and DrugBank. The structural overlap was performed using the InChIKey representation of the compound in SureChEMBL against the three resources.

Upon identifying common compounds across these resources (Fig. 3a), two key findings were revealed. Firstly, only 0.02% (2,096 out of 10 million) of the patent compounds were eventually approved for one or more indication areas, according to data extracted from DrugBank, and secondly, PubChem retrieved compounds exhibited the highest overlap (91.5%) with the patent compounds from SureChEMBL. In contrast, ChEMBL demonstrated only a 0.1% overlap with patent compounds in SureChEMBL, an indicator that both resources occupy different chemical spaces. As illustrated in Fig. 3a, a small percentage (5.5%) of patent compounds were specific to the SureChEMBL database. Among these compounds, more than half were mined from US-based patent documents, while the remaining have been mined from EPO- or WIPO-based patent documents. Additionally, an examination of the annual count of SureChEMBL-specific compounds revealed a gradual decrease over time (Fig. 3b).

Images as the major source for compound annotation in patent documents. A patent document is a structured document containing sections including the title, abstract, description and claims[38]. Among these patent document sections, determining the location from which a compound was mined can provide insights into the correlation between the compound and the patent's applicability. For instance, if a compound was mentioned in the description section, it is likely to be associated with prior art (i.e. "referenced" compounds) relevant to the patent document. On the other hand, a compound mentioned in the claims section would likely pertain to the novel invention disclosed in the patent document.

In SureChEMBL, a patent document consists of four sections: title, abstract, description and claims. Along with these specific sections, chemical structure images and molfile (specifically restricted to patent documents collected from USPTO) serve as sources of compound annotation in SureChEMBL. Notably, these later sources (i.e. images and molfiles) were only annotated for patent applications after 2007[27]. Together, these six

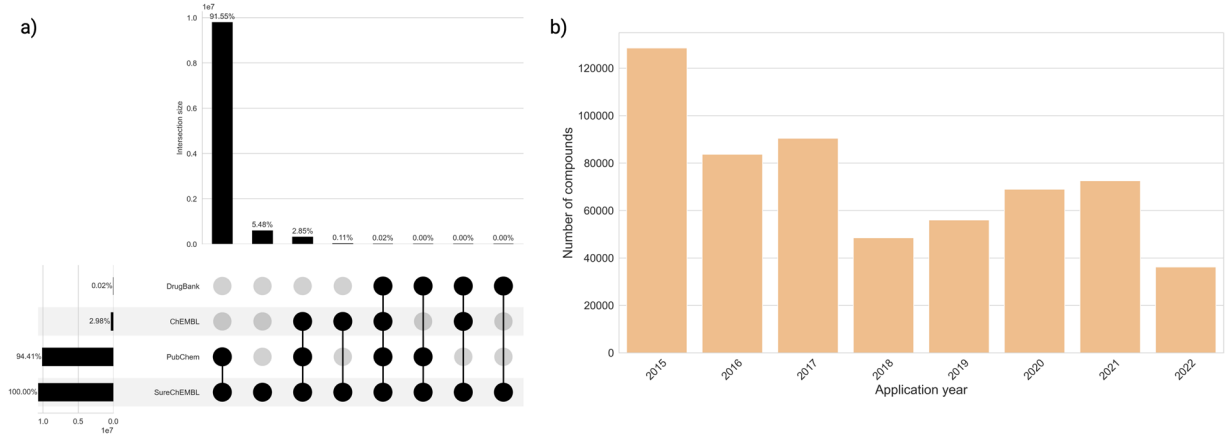

Fig. 3 (**a**) Distribution of patent compounds across four chemical resources, namely SureChEMBL, PubChem, ChEMBL and DrugBank. (**b**) Distribution of the proportion of patent compounds specific to SureChEMBL. The figure was formatted with BioRender.com.

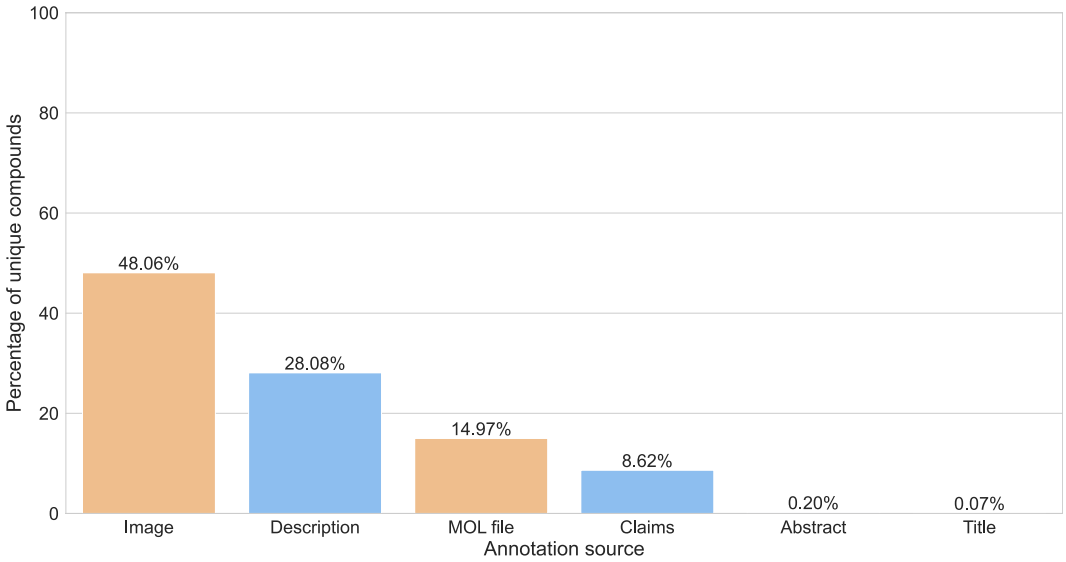

Fig. 4 Percentage of compounds annotated from the various patent document sources. Each bar in the figure corresponds to deduplicated compounds annotated specifically from the individual section of the patent document. The textual sections of a patent document (blue) are distinguished from the additional sources for annotation (orange) based on their colour.

sources of the patent document were utilized for biomedical entity annotations in SureChEMBL using numerous public and proprietary mining tools[27]. Hence, to provide an overview of the major sources surrounding the annotation of patent compounds, we investigated the sections frequently mined and annotated for compounds in SureChEMBL. We first calculated the average number of sources associated with patent documents in SureChEMBL. This analysis revealed that approximately 31.2% of patent compounds are found in more than one of the six sources. Next, we performed a thorough examination of the sources with regard to the compounds. We found that the description section (with ~28.08%) of the patent document was the major source of textual data involved in the extraction of patent compounds (Fig. 4). As illustrated in the figure, both the additional patent document sources (images with ~48% and molfiles with ~15%) were part of the top three sources for data annotations in SureChEMBL.

Over half of patent compounds show compliance with Ro5 framework. To improve the efficiency of the drug discovery process, scientists have formulated guidelines or rules based on key determinant properties of compounds of drug likeness. Lipinski[39] and Veber *et al.*[40] provided the framework for the Rule of Five (Ro5), depending on physicochemical properties, to enhance the oral bioavailability of a compound[39–41]. Later, Doak *et al.* (2014), along with other researchers, extended the Ro5 for the oral bioavailability space of drugs, referring to it as the beyond Ro5 (bRo5) space[42–44]. The criteria of bRo5 supported the selection of cell-permeable

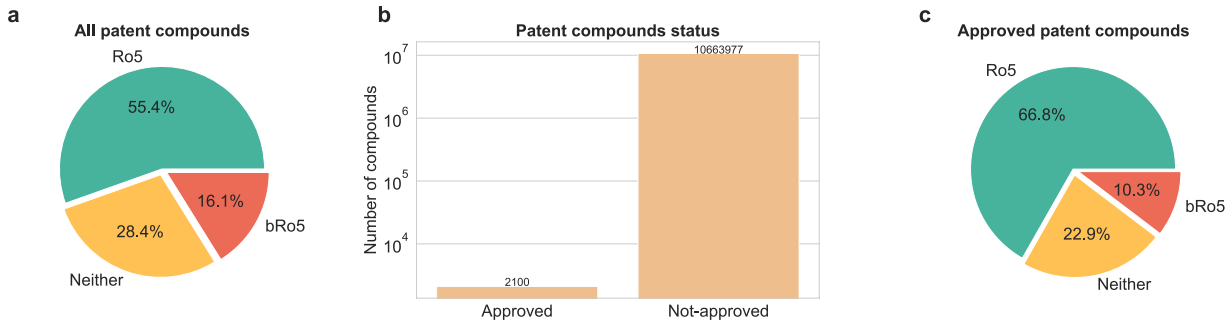

Fig. 5 Drug-like compliance of patent compounds. (**a**) Radial chart demonstrating the percentages of the Ro5 and bRo5 framework compliances of patent compounds. As shown, about 30% of patent compounds do not comply with either of the two frameworks. (**b**) Overview of the drug approval status of patent compounds. (**c**) Radial chart demonstrating the percentages of the Ro5 and bRo5 framework compliances of patent compounds that are approved drugs (i.e. 2,100 compounds).

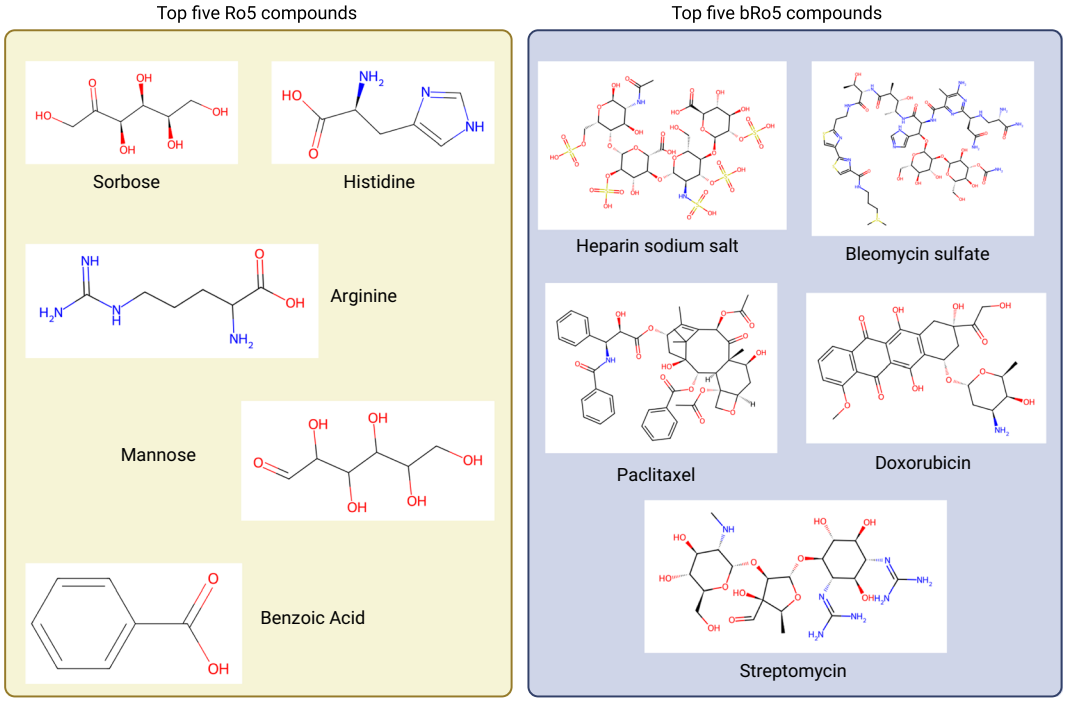

Fig. 6 Top five prevalent patent compounds in the Ro5 and bRo5 categories.

clinical candidates that demonstrated good pharmacokinetics (PK) and explored the "undruggable" targets, both of which could not have been possible previously with the Ro5 filtering.

To profile the drug-like space for the patent compounds, we explored the Ro5 and bRo5 space of these compounds. This analysis revealed that in the past decade, 55.46% of compounds complied with the Ro5 framework, and 16.11% compounds did so with the bRo5 (Fig. 5). The remaining compounds (28.43%) complied neither with the Ro5 nor bRo5 spaces. Next, we divided the compounds into two categories, approved and non-approved, based on the data in DrugBank. As mentioned in the previous sections, a very low number of patent compounds were approved. Moreover, a consistent trend was found, with more than half of the compounds complying with the Ro5 framework, among both approved and non-approved (Fig. 5).

Finally, we explored the most frequent patent compounds, focusing on the top five based on their prevalence in patent documents (Fig. 6). In the Ro5 class, traditional sugars such as Sorbose (SCHEMBL762) and Mannose (SCHEMBL1812) were found in over 200,000 and 150,000 patent documents, respectively. Essential amino acids like histidine (SCHEMBL3259) and arginine (SCHEMBL1790) were also prevalent, appearing in 150,000 to 200,000 patent documents. On the other hand, in the bRo5 class, we found heparin (SCHEMBL11557), a naturally occurring human metabolite, in more than 95,000 patent documents. Moreover, prominent drugs such as Paclitaxel (SCHEMBL3976), Bleomycin sulphate (SCHEMBL1599) and Doxorubicin (SCHEMBL3243), which are therapeutic drugs for treating cancer and antibacterial drug Streptomycin (SCHEMBL3276) were other patent compounds found in the bRo5 class with occurrence in about 90,000 patent documents. It is important

Year	Molecular weight	LogP	# HBA	# HBD	# RotB	# Rings	# Stereoisomers
2015	407.84	3.92	5.07	1.49	5.92	3.46	4.40
2016	437.60	4.29	5.36	1.55	6.31	3.84	5.05
2017	450.68	4.56	5.33	1.62	6.52	4.01	7.01
2018	451.07	4.61	5.35	1.53	6.52	4.02	6.57
2019	460.80	4.67	5.49	1.57	6.48	4.21	6.52
2020	471.94	4.90	5.56	1.55	6.52	4.42	6.47
2021	474.90	4.90	5.58	1.57	6.51	4.47	5.78
2022	498.41	5.43	5.70	1.50	6.71	4.94	6.37

Table 2. Physicochemical properties of patent compounds. For compounds found in each year, an average of the different 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 (# RotB), (vi) the number of any ring (# Rings) and (vii) the number of stereoisomers were calculated.

to acknowledge that when dealing with thousands of patent documents associated with a compound, not all of them would be irrelevant. In fact, for a successfully approved active pharmaceutical ingredient (API) with potential, many follow-up patent applications may emerge. These could pertain to its synthesis, specific drug delivery system, novel indication area, or potential combination therapy with another ingredient. Moreover, APIs are frequently cited as prior art in patent documents, underscoring their significance in the pharmaceutical landscape.

Patent compounds show signs of enhanced chemical structural diversity. Recently, PROteolysis-TArgeting Chimeras (PROTACS) have been identified as novel therapeutics with the potential to progress into clinics[45,46]. They achieve protein degradation by "hijacking" the cell's ubiquitin-proteasome system (UPS) and bringing together the target protein and an E3 ligase. Due to their non-adherent Ro5 characteristics, these molecules have not undergone "classical" prior optimization for oral bioavailability[47] and CNS penetration[48]. 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 functional groups or side chains, allowing for ring-shaped structures[49,50]. With the increasing interest in such compounds as potential clinical and drug candidates, we determined the physicochemical properties of patent compounds to check whether the chemical space expansion reflected PROTAC-like and macrocyclic compounds, among others, in recent years.

Table 2 summarizes the average physicochemical characteristics of patent compounds found annually. A gradual increase in characteristic molecular properties (i.e. molecular weight, the hydrogen bond donor and acceptors, LogP, and the number of rings) was observed. These properties, especially molecular weight, were nearing the upper limit of the Ro5 criteria. Specifically, the average molecular weight for patent compounds was between 400 and 500 Daltons, the average LogP was between 4 and 5, the average number of hydrogen bond acceptors and donors were below 6 and 2, respectively, and the average number of rotatable bonds was between 6 and 7. In addition, a consistent increase in the number of rings in patent compounds was also found in the past decade. Our analysis also uncovered some common findings with respect to molecular properties across potential clinical candidates. For example, we observed that over a million patent compounds exhibited with two or more chiral centers. This proportion supasses the number of compounds with only one chiral center. For either group, no enantiopurity is registered, so it is not possible to roughly assess how many enantioselective synthetic steps have been used. This finding aligns closely with recent research conducted by Scott *et al.*[51].

Additionally, we were interested in identifying bioactive compounds that could be covalent binders, showed existing polypharmacology, or were reactive in nature. To achieve this, one strategy involved examining the published biological activities and selectivities of the compounds. However, adopting this approach could lead to a very small subset of patent compounds (2,000–5,000), potentially yielding inconclusive results due to the existence of data silos surrounding the publication of biological data in patent documents. Thus, we used an alternative approach to understand the polypharmacological nature of patent compounds. This involved confirming the presence of Pan-Assay INterference Structures (PAINS), which are frequently used in drug discovery to flag and mark compounds that could cause interference in bioassays[44,52]. Consequently, such compounds that contain one or more PAINS alerts are usually removed during pre-clinical research due to their polypharmacological nature marking them as high risks for off-target effects.

In total, we identified 277 PAINS alerts in the patent compounds. This represents approximately 3.7% of all patent compounds in SureChEMBL that show the presence of at least one of these PAINS alerts (Fig. 7a). The most prominent of these PAINS alerts include the presence of azo compounds (Azo_a(324); 18.7%; 71,014 compounds), the presence of compounds involving one aniline and two alkyl groups with either an additional alkyl group (Anil_di_alk_a(478), 14%; 53,127 compounds) or an additional carbon (Anil_di_alk_c(246), 5.7%; 21,662 compounds), the presence of compound containing an indole, a phenyl and an alkyl group (Indol_3yl_alk(461), 8.5%; 32,290 compounds), and the presence of compounds containing catechol substructures (Catechol_a(92), 5.9%; 22,452 compounds). Moreover, aromatic PAINS like quinone also make it to the top of the list with about 5.2% (19,665) patent compounds. Interestingly, bioassay reagent materials like dyes (15,568 compounds) and Mannich bases (15,154 compounds) also appear in this list of PAINS alerts.

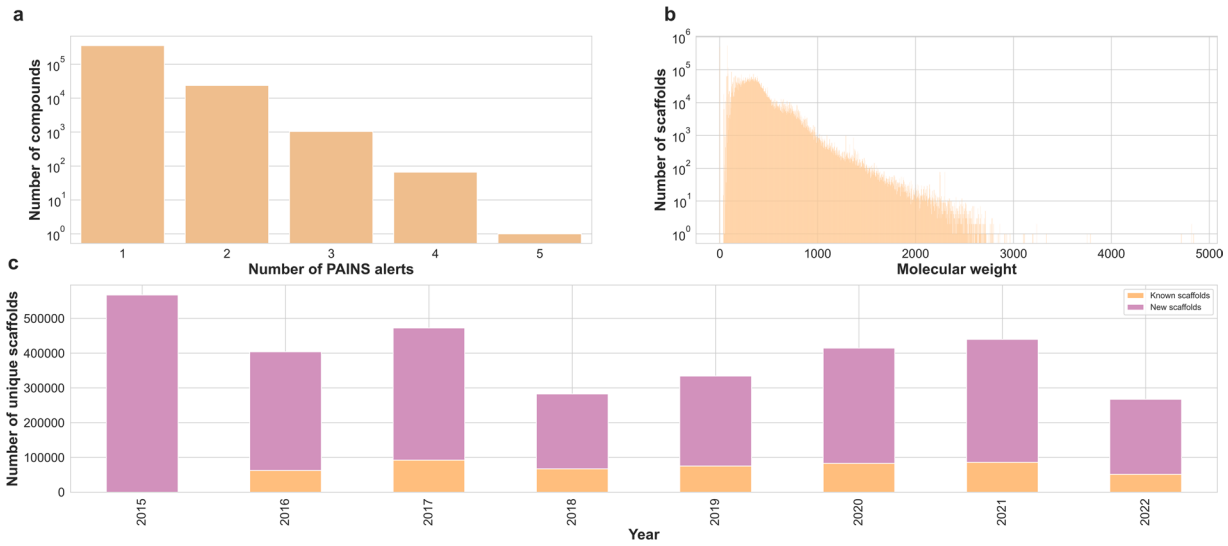

Fig. 7 (**a**) Count of PAINS alerts found to be associated with patent compounds. (**b**) Distribution of the molecular weight of the generic Murcko scaffold across the patent compounds. (**c**) Count of unique Murcko scaffolds found per year in patent documents. For each year, we distinguish the "known" scaffolds (orange) from "new" scaffolds (pink) based on the occurrence of the Murcko scaffold SMILES in previous years.

Patent Field	Number of scaffolds	Percentage of scaffold (%)
Abstract	8,120	0.19
Claims	315,413	7.73
Description	811,216	19.89
Image (after 2007)	2,261,009	55.44
MOL file (after 2007)	677,947	16.62
Title	4,494	0.11

Table 3. Summary of the number of unique scaffolds found in the individual patent document sections.

To conclude this exploration, we reduced the compounds to their generic Bemis-Murcko (BM) scaffold to quantify the scaffold diversity of patent compounds. 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 chemical definition would provide clues on how they were derived, and second, this representation retains the rings and side chains found in the compounds, unlike the graph framework that replaces all heteroatoms to carbon and collapses all bonds to single bonds notations[53]. This analysis revealed that 3 million distinct scaffolds encompass patent compounds in SureChEMBL. These scaffolds cover a broad range of molecular sizes, spanning from a compact molecule of 38.01 Daltons to a large molecule of 4841.19 Daltons (Fig. 7b). The year-wise comparison of the BM scaffold revealed that annually an average of 332,942 new generic scaffolds were patented (Fig. 7c). Moreover, trends showcasing a fluctuating number of scaffolds with a recent decline around the COVID-19 pandemic (2021–22) were seen. Tracing back the patent document source (i.e., the section or source from which the compound was annotated) revealed that more than half (55.44%) of the scaffolds were found to be associated with chemical images in patent documents, while only 19.89% were associated with the description section of patent document. Moreover, about 16.62% were found to be extracted from the claims section of the patent document (Table 3).

In this analysis, it is necessary to acknowledge that reducing the compounds to their generic BM scaffolds would result in the generation of promiscuous 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. Hence, our approach also unveiled a larger number of such common scaffolds. Figure 8 summarizes the top ten commonly identified scaffolds in patent compounds. As shown, the majority of these scaffolds have a single ring with heteroatoms causing it to weigh about a few hundred daltons. The most prominent ones include single cyclic scaffolds such as cyclohexane and tetrahydropyran or double-ring compounds like naphthalene and diphenylmethane.

Tracing a subset of approved drugs back to their patent documents. Pre-established regulations like the Patent Act of 1990 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[54,55]. Consequently, many disparities have arisen between the drug names present in patent documents

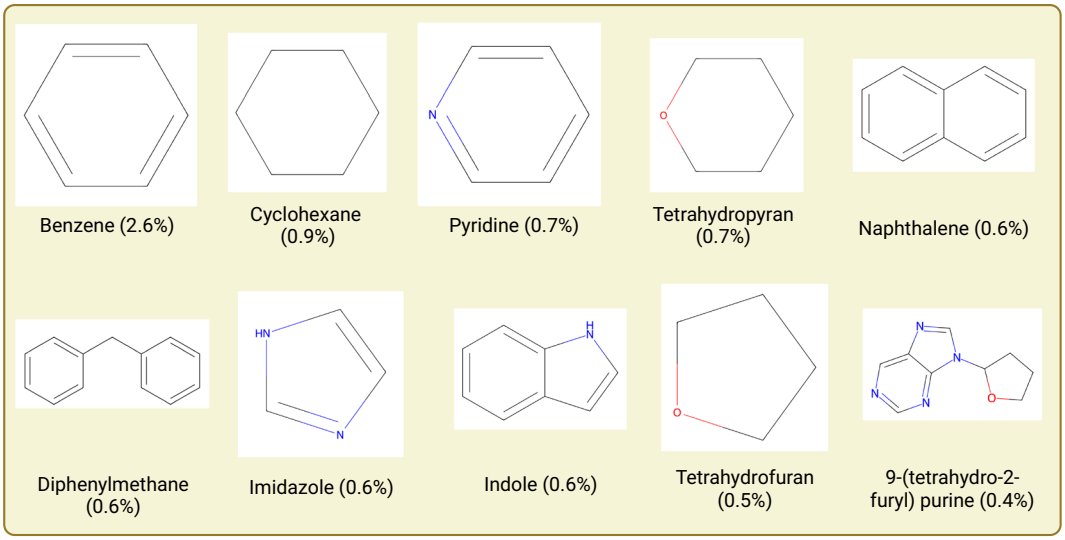

Fig. 8 The top Murcko scaffold of compounds and their respective frequency found in patent documents. The percentages only sum up to around 8.2% of the total patent compound scaffolds, a clear sign of the structural diversity within the patent chemical space.

and their corresponding brand name, posing a challenge in finding associated patent documents[56]. In the past, successful endeavours 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[57]. Acknowledging the complexity of the drug nomenclature across the different stages of drug development, we leveraged the chemical representation (InChIKey) of patent compounds to generate an inventory of their corresponding clinical status in humans.

To do so, we started by looking into the intersection of the chemical space of patent compounds with investigational (i.e. drugs that have reached clinical trials) and withdrawn (i.e. drugs that have been discontinued) drugs in DrugBank. We found that only 3,235 of the ten million patent compounds have reached clinical trials, with a mere 0.0008% (85 compounds) falling under the withdrawn drugs category (Fig. 9a). In addition, databases such as ChEMBL enable identifying the drug research status (i.e., preclinical, clinical, and approved) of compounds, and hence we leverage this resource to identify the farthest research stage a patent compound traversed to in the drug discovery pipeline. Figure 9b depicts that compounds in patent documents are distributed across various research phases, ranging from preclinical to clinical, with the majority concentrated in the preclinical stage. Of these patent compounds, roughly 1% of the drugs were approved for one or more indication areas, according to ChEMBL. Furthermore, 1.6% of the patent compounds had no information (classified as "unknown" by ChEMBL) regarding their clinical stage and were likely to be lost during translation from patent documents to clinics or failed to be captured and annotated by the resource database. Furthermore, Phases 2 and 3 of clinical trials exhibited a greater proportion of patent compounds than Phase 1.

Discussion

Patent documents play an essential role in drug discovery and biological annotation pipelines, such as those that capture a molecule's image and convert it to SMILES, or those that highlight gene and disease names in the patent documents. This work focuses on patent compounds found in SureChEMBL, a patent database for life sciences, and assesses the annotation quality for drug-like molecules and drug discovery-related documents. To the best of our knowledge, this is the first systematic effort done in this direction with the entire database.

Initially, we started by inspecting the jurisdiction coverage of patent documents in SureChEMBL. As expected, countries such as the United States and Europe had the highest number of patent compounds. In contrast, a very low percentage of patent documents from Japan (through JPO) were present. This low number is confirmed by SureChEMBL, acknowledging their limited access to bibliographic information from Japanese patent documents (i.e. titles and abstracts) provided by the JPO[27,58]. Furthermore, challenges arise in converting Japanese text to English for the ingestion and storage of biomedical entities in SureChEMBL, thereby exacerbating the limitations. Previously, the absence of machine-translated full texts from which patent compounds could be extracted posed a bottleneck. However, recent advancements in integrated AI annotation within the resource offer potential mitigation for this issue in the future (https://www.ebi.ac.uk/about/news/technology-and-innovation/ai-annotations-increase-patent-data-in-surechembl/). Next, we looked into the quantitative aspect of data in SureChEMBL. A small fraction (0.2%) of compounds were identified in more than one patent document. This could entail the presence of either repurposing patent documents (i.e., a patent application on the reuse of known drugs for a different indication) or compounds annotated from utility patent documents (i.e., a patent application dealing with approved drugs in different pharmaceutical forms or route of administration for specific treatments).

In order to assess the ease with which patent compounds can be identified in the literature, we searched the compound structures across three major compound databases (i.e. PubChem, ChEMBL and DrugBank).

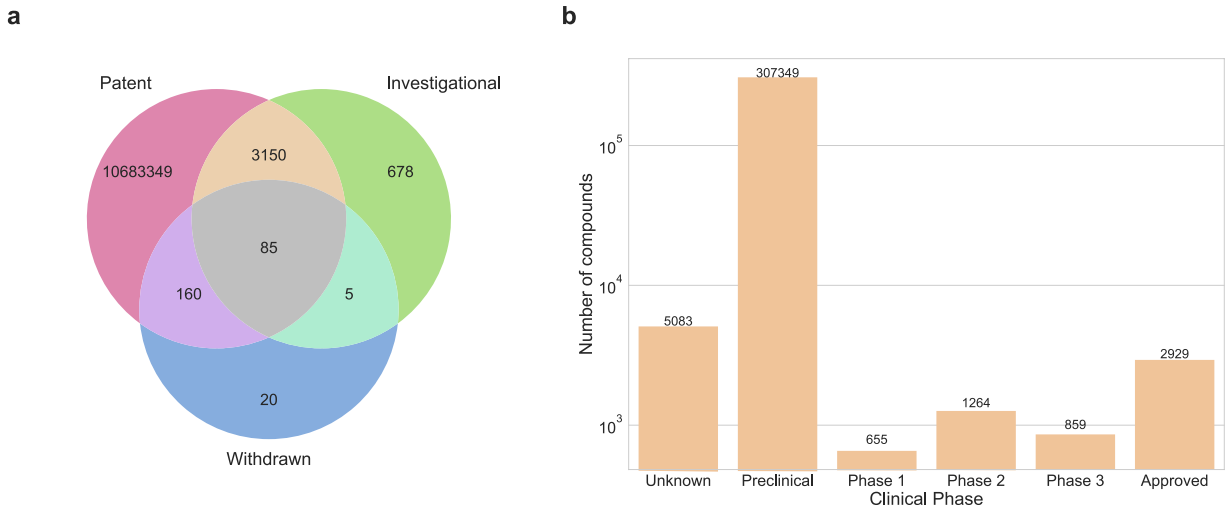

Fig. 9 (**a**) Euler figure of the chemical space across the patent compounds, investigational and withdrawn drugs found in DrugBank. (**b**) Distribution of patent compounds in the different clinical phases as per ChEMBL.

In this analysis, we found that a very low percentage of approved drugs (found in DrugBank) were a part of SureChEMBL. This low percentage could be attributed to three reasons: Firstly, patent documents are often crafted to encompass a broader chemical structure-activity landscape than the clinical candidates, thus safeguarding the candidates' secrecy[19,47]. This is typically achieved through the use of Markush structures in patent claims, allowing for coverage of a wide range of structural variations, including potential drug candidates that may still be unknown at the time of patent filing. Additionally, active pharmaceutical ingredients (APIs) are utilized to extend the applicability of the patent. Secondly, the changes in the nomenclature of the drug candidate as it progresses through the clinical pipeline, obscure its visibility. Despite SureChEMBL's cross-reference dictionary enabling the retrieval of patent compounds through multiple depictions (e.g., SMILES, IUPAC names, etc.), this limitation arises from inconsistencies in the names used by patent assignees or holders. Additionally, this could be due to the limited information provided by the patent assignees or holders in the patent document (a consequence of the former reason), thus hindering automated systems and pipelines (like SureChEMBL) from accurately recognizing the exact structure of the approved drug. Finally, the third limitation pertains to the use of DrugBank as a proxy for representing the approved drug space. DrugBank, being a commercial database, provides limited information for academic research. Moreover, our results also revealed an analogous chemical space being occupied between patent compounds in SureChEMBL and PubChem. This is unsurprising, given that PubChem leverages the SureChEMBL database to broaden its underlying chemical space. Furthermore, it's worth noting that in the near future, PubChem could accommodate additional patent compounds through its automated patent annotation pipeline. This pipeline establishes connections between compounds and relevant patent documents referenced in Google Patents[59]. Furthermore, a small proportion of compounds were not found in any public compound resources and were instead confined to SureChEMBL, indicating the presence of compounds from proprietary libraries used by patent assignees for drug discovery.

Following this, we investigated the major sources for compound annotation within SureChEMBL. These sources included known sections of patent documents (i.e. title, abstract, description and claims), images and MOL files. The description section was identified as the prominent textual source for compound annotation, highlighting that the annotated compound could be a part of the primary invention, whether it be related to its synthesis, formulation or application within a specific area. Moreover, it is essential to note that the text within the description section could also be too general, and 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. Moreover, a small proportion (~15%) of compounds are also annotated from the MOL file, which is one of the basic files required for compound patent documents filed to the USPTO.

Additionally, we explored the chemical space of patent documents to delineate their structural diversity and drug-likeness space. Naturally, most of the compounds complied with the Ro5 framework for drug-likeness. This was further confirmed by looking at the underlying physicochemical distribution of patent compounds (as shown in Table 2). We also identified a small proportion of patent compounds outside the Ro5 space, falling into the bRo5 space; a similar trend was observed in the case of approved drugs in recent years[43]. Besides this, an increase in the physicochemical properties of the patent compounds was observed, which may have been due to multiple factors, including increasing interest in the development of PROTAC-like and signalling macrocyclic compounds (for e.g., cyclic peptides or cyclic kinase inhibitors)[60] or progress with regards to chemical synthesis capabilities[61,62]. Our exploration regarding the bioactivity of patent compounds led to the recognition of PAINS liable compounds in SureChEMBL. The presence of these assay-interfering compounds is not surprising given that a study by Capuzzi *et al.*[63] showed that FDA-approved drugs containing PAINS were more active than non-PAINS-containing drugs[63]. This has indeed led researchers like Senger *et al.*[64] to question the filtering of promiscuous compounds during the early drug discovery steps[64]. Nevertheless, the question of whether these

PAINS alert patent compounds are problematic (showing false positive results in assays) or innocuous (due to their possible polypharmacology activity) remains unsolved. From our perspective, this finding highlights the notion that the mere identification of compounds from patent documents is not sufficient to identify potential drug candidates. In the case of SureChEMBL, there is a need for performing a medicinal chemistry-oriented filtration to eliminate non-active or activity-interfering patent compounds prior to their downstream utility.

In a similar manner, we also explored the generic BM scaffolds of the patent compounds. This exploration showed a decline in the number of scaffolds generated in the past years. This could be attributed to several reasons, such as exceptional factors like the COVID-19-dependent blockade of some patenting activities in 2022, or structural reasons rooted within medicinal chemistry syntheses pipelines. Lastly, we concluded our analysis by addressing the drug discovery path of a patent drug by looking into its transition from a patent document to post-approval. Here we reported that of all the patent compounds found in ChEMBL, only 1% of the compounds were approved drugs. This is not surprising as an analysis by Brown[65] showed that hit compounds evolve through the drug discovery pipeline as they undergo structural modification that ensures their entry into clinics[65]. Moreover, a larger proportion of patent compounds were found in Phase 2 and Phase 3 than in Phase 1. This is attributed to the fact that these trial phases (i.e. Phases 2 and 3) typically yield a larger number of scientific publications, assuming the trials were successful[66]. In conclusion, our work provides a medicinal chemistry perspective on the chemical landscape formed by patent compounds, thus laying a foundation for the future utility of SureChEMBL. Furthermore, we believe that understanding the state of the art in terms of patent compounds and their scaffolds is crucial for enhancing innovation by exploring novel chemical space while minimizing the risks associated with inadvertently reusing chemical space for specific and undesired target indications.

We acknowledge certain limitations in our analysis that warrant attention. 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, as discussed in our study. Secondly, we assumed that all patent compounds in SureChEMBL are relevant to drug discovery, as they may pertain to either an indication area or drug formulation. However, this assumption may not always hold true, and it would be preferable to refine our analysis by focusing on a subset of patent IPC codes, as outlined by Gadiya *et al.*[67], to create a more drug discovery-specific patent document dataset. Also, the annotation source in SureChEMBL does not include the context from which the compounds were annotated in the patent document, at least in its data dump. This shortcoming makes it difficult to distinguish compounds that have been "referred" (i.e. prior art) in patent documents from those "claimed" (i.e. novelty). This has eventually led researchers to perform an additional layer of annotation on top of SureChEMBL-extracted patent documents[18,20]. Lastly, we briefly look into drug repurposing patent documents and recognize the potential value in identifying specific indication areas where drug repurposing patent documents are concentrated. This could offer insights into the similarity of MoA across different disease indications for certain compounds.

Discussions on the efficiency of patent documents for drug discovery have been raised numerous times, given their inability to disclose the drug candidate, thereby protecting the compounds' IP. This study aims to shed light on the utility of compound data in patent documents in the context of drug discovery. By leveraging SureChEMBL as the patent resource and untapping its chemical space, we open the avenue for the use of this resource for chemoinformatic-based models. For example, patent compounds could be used to extend and expand existing chemical datasets by enriching the structure-activity relationship landscape around the lead candidate. Such an approach could be a potential alternative to the generative AI-based approaches, provided that a patent document around the lead molecule exists and has been previously mapped. Hence, SureChEMBL has vast potential that has yet to be mined and leveraged for drug discovery purposes.

Methods

Collecting compound data from patent literature. We used SureChEMBL (https://www.surechembl.org/), an extensive publicly available patent compound data catalogue, as a source for patent documents and metadata[27]. We obtained all the tab-separated data files (.txt) from the FTP server of the resource (ftp://ftp.ebi.ac.uk/pub/databases/chembl/SureChEMBL/data/map/). The legacy data from 1994–2014 (identified by the file name SureChEMBL_map_20141231.txt.gz) had a different data format, lacking patent information, which complicated the transition from compounds to corresponding patent documents. Consequently, the legacy data was excluded from the analysis presented in the study. Ultimately, the compounds from 2015–2022 were utilized as the patent compound collection for this work.

Compound databases utilized for compound metadata annotation. To identify compounds annotated within SureChEMBL, we mapped them to three large independent compound data resources, namely PubChem (v.2023)[59], ChEMBL (v.32)[68], and DrugBank (v.5.1.10)[69]. The data from these resources was obtained either by querying the REST API service (as in the case of PubChem through PubchemPy), or by downloading a local data dump of the resources (as in the cases of DrugBank, via an academic licence, and ChEMBL, via the SQL database from their FTP server). A compound is said to be the same across two resources provided that an exact match of the InChIKey is present.

Moreover, we leveraged the clinical stage annotation of compounds in ChEMBL ("max phase") to annotate the clinical phase of a corresponding compound found in SureChEMBL. Additionally, DrugBank was used to validate the compounds that have been approved and distinguish them from those that had been withdrawn. It is worth noting that the commercial nature of DrugBank, while ensuring faster updates than public databases, might limit certain information to premium users.

Chemical space exploration using physicochemical properties of compounds. The chemical space of compounds was explored using two well-defined and established rules:

a) Ro5 - Extending Lipinski's Rule of Five (Ro5)[39,40], Veber *et al.*[40] added rotatable bonds (NRotB) and the topological polar surface area (TPSA) as additional features, proposing that compounds should have a TPSA < 140 and NRotB < 12 to enhance the probability of sufficient oral bioavailability[42].
b) beyond Ro5/bRo5 - According to Doak *et al.* (2016), compounds with $500 <$ MW < 3000 daltons with at least one property beyond the *extended* Ro5 (i.e., LogP > 7.5 or LogP < 0, hydrogen bond donors (HBD) > 5, hydrogen bond acceptors (HBA) > 10, polar surface (TPSA) > 200, and rotatable bonds (NRotB) > 20 fall in this category[43].

A desalting step using RDKit was performed for the patent compound. In addition to these two rules, we also assessed the chemical space underlying patent documents by conducting a scaffold diversity assessment on compounds derived from patent documents. To accomplish this, we simplified the desalted compounds into their Murcko scaffolds, preserving the generic forms of ring components, linkers and side chains[44]. These Murcko scaffolds were then used to examine the occurrence of newly registered scaffolds on an annual basis. The generic Murcko scaffold for the patent compounds was generated using RDKit's "Scaffolds.MurckoScaffold.MurckoScaffoldSmiles()" function. We also identified PAINS alerts within the patent compounds. This was performed using the RDKit library using the codebase from the TeachOpenCADD tutorials[70].

Data availability

The data used in this study can be accessed on Zenodo[71]. The "Figures" directory consists of all the figures shown in this study. The "Mappings" directory consists of JSON serialized data dumps corresponding to the physiochemical properties, PAINS alerts and Murcko scaffolds for compounds found in SureChEMBL. The "Processed" directory involved the successful mapping of SureChEMBL compounds to external public databases like PubChem and ChEMBL. Finally, the "Raw" directory includes the combined original data dump of SureChEMBL from its FTP server (ftp://ftp.ebi.ac.uk/pub/databases/chembl/SureChEMBL/data/map/).

Code availability

The Python scripts and Jupyter notebooks supporting the conclusions of this study can be accessed and downloaded via GitHub (https://github.com/Fraunhofer-ITMP/patent-clinical-candidate-characteristics). The repository is structured into "Data" and "Notebook" sections. The "Data" section is the exact replica of the Zenodo dump mentioned previously. The "Notebook" section includes all the analyses presented in this study organized based on the sections within the results.

Received: 25 January 2024; Accepted: 13 May 2024;
Published online: 16 May 2024

References

1. Grabowski, H. G., DiMasi, J. A. & Long, G. The roles of patents and research and development incentives in biopharmaceutical innovation. *Health Affairs* **34**, 302–310 (2015).
2. Kesselheim, A. S., Sinha, M. S. & Avorn, J. Determinants of market exclusivity for prescription drugs in the United States. *JAMA Internal Medicine* **177**, 1658 (2017).
3. Dunn, M. K. Timing of patent filing and market exclusivity. *Nature Reviews. Drug Discover/Nature Reviews. Drug Discovery* **10**, 487–488 (2011).
4. Sayle, R. A., Petrov, P., Winter, J. & Mureşan, S. Improved chemical text mining of patents using infinite dictionaries, translation and automatic spelling correction. *Journal of Cheminformatics* **3** (2011).
5. Gadiya, Y., Gribbon, P., Hofmann-Apitius, M. & Zaliani, A. Pharmaceutical patent landscaping: A novel approach to understand patents from the drug discovery perspective. *Artificial Intelligence in the Life Sciences* **3**, 100069 (2023).
6. Kong, X. *et al.* STING as an emerging therapeutic target for drug discovery: Perspectives from the global patent landscape. *Journal of Advanced Research* **44**, 119–133 (2023).
7. Zhang, H. & Li, Y. The patent landscape of BRAF Target and KRAS Target. *Recent Patents on Anti-cancer Drug Discovery* **18**, 495–505 (2023).
8. Song, C. H., Han, J., Jeong, B. & Yoon, J. Mapping the patent landscape in the field of personalized medicine. *Journal of Pharmaceutical Innovation* **12**, 238–248 (2017).
9. Lahiry, S. R. & Rangarajan, K. Patent landscape for Indian biopharmaceutical sector: A Strategic insight. in *Flexible systems management* 31–47, https://doi.org/10.1007/978-981-10-8926-8_3 (2018).
10. Mucke, H. A. Intellectual property considerations. in *The Royal Society of Chemistry eBooks* 264–279, https://doi.org/10.1039/9781839163401-00264 (2022).
11. Strittmatter, S. M. Overcoming Drug Development Bottlenecks With Repurposing: Old drugs learn new tricks. *Nature Medicine* **20**, 590–591 (2014).
12. Senger, S. Assessment of the significance of patent-derived information for the early identification of compound–target interaction hypotheses. *Journal of Cheminformatics* **9** (2017).
13. Colen, L., Belderbos, R., Kelchtermans, S. & Leten, B. Many are called, few are chosen: the role of science in drug development decisions. *The Journal of Technology Transfer* https://doi.org/10.1007/s10961-022-09982-6 (2023).
14. Schmitt, V. J., Walter, L. & Schnittker, F. C. Assessment of patentability by means of semantic patent analysis – A mathematical-logical approach. *World Patent Information* **73**, 102182 (2023).
15. Fabry, B., Ernst, H., Langholz, J. & Koster, M. P. 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**, 215–225 (2006).
16. Grego, T., Pęzik, P., Couto, F. M. & Rebholz-Schuhmann, D. Identification of chemical entities in patent documents. in *Lecture notes in computer science* 942–949, https://doi.org/10.1007/978-3-642-02481-8_144 (2009).
17. Farre-Mensa, J., Hegde, D. & Ljungqvist, A. What Is a Patent Worth? Evidence from the U.S. Patent "Lottery". *The Journal of Finance* **75**, 639–682 (2019).
18. Falaguera, M. J. & Mestres, J. Identification of the core chemical structure in SUReCHEMBL patents. *Journal of Chemical Information and Modeling* **61**, 2241–2247 (2021).

19. Falaguera, M. J. & Mestres, J. Congenericity of claimed compounds in patent applications. *Molecules* **26**, 5253 (2021).
20. Kunimoto, R. & Bajorath, J. Exploring sets of molecules from patents and relationships to other active compounds in chemical space networks. *Journal of Computer-aided Molecular Design* **31**, 779–788 (2017).
21. Wagner, Ş., Sternitzke, C. & Walter, S. G. Mapping Markush. *Research Policy* **51**, 104597 (2022).
22. Deng, W., Berthel, S. J. & So, W. V. Intuitive patent Markush Structure Visualization tool for medicinal chemists. *Journal of Chemical Information and Modeling* **51**, 511–520 (2011).
23. Wills, T. J. & Lipkus, A. H. Structural approach to assessing the innovativeness of new drugs finds accelerating rate of innovation. *ACS Medicinal Chemistry Letters* **11**, 2114–2119 (2020).
24. Kim, J. & Lee, S. Patent databases for innovation studies: A comparative analysis of USPTO, EPO, JPO and KIPO. *Technological Forecasting & Social Change* **92**, 332–345 (2015).
25. Marco, A. C., Graham, S. & Apple, K. The USPTO Patent Assignment Dataset: Descriptions and Analysis. *Social Science Research Network* https://doi.org/10.2139/ssrn.2849634 (2015).
26. Hill, L. L. The Orange Book. *Nature Reviews. Drug Discovery* **4**, 621 (2005).
27. Papadatos, G. *et al*. SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Research* **44**, D1220–D1228 (2015).
28. Ferrence, G. M. *et al*. CSD Communications of the Cambridge Structural Database. *IUCrJ* **10**, 6–15 (2023).
29. Southan, C., Sitzmann, M. & Mureşan, S. Comparing the chemical structure and protein content of CHEMBL, DrugBank, Human Metabolome Database and the Therapeutic Target database. *Molecular Informatics* **32**, 881–897 (2013).
30. Ghani, S. S. A comprehensive review of database resources in chemistry. *Eclética Química* **45**, 57–68 (2020).
31. Tamura, S., Miyao, T. & Bajorath, J. Large-scale prediction of activity cliffs using machine and deep learning methods of increasing complexity. *Journal of Cheminformatics* **15** (2023).
32. Van Tran, T. T., Wibowo, A., Tayara, H. & Chong, K. T. Artificial intelligence in Drug toxicity Prediction: Recent advances, challenges, and future perspectives. *Journal of Chemical Information and Modeling* **63**, 2628–2643 (2023).
33. Lagunin, A. *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**, 1689 (2023).
34. Chen, W., Liu, X., Zhang, S. & Chen, S. Artificial intelligence for drug discovery: Resources, methods, and applications. *Molecular Therapy. Nucleic Acids* **31**, 691–702 (2023).
35. Bhattacharjee, A. K. Pharmacophore-based virtual screening of large compound databases can aid "big data" problems in drug discovery. in *Elsevier eBooks* 231–246, https://doi.org/10.1016/b978-0-323-85713-0.00014-1 (2023).
36. Almansour, N. M., Allemailem, K. S., Aty, A. A. A. E., Ismail, E. I. F. & Ibrahim, M. A. A. In Silico Mining of Natural Products Atlas (NPATLAS) database for identifying effective BCL-2 inhibitors: molecular docking, molecular dynamics, and pharmacokinetics characteristics. *Molecules* **28**, 783 (2023).
37. Ohms, J. Validity of PubChem compounds supplied by Patentscope or SureChEMBL. *World Patent Information* **70**, 102134 (2022).
38. Jessop, D., Adams, S. & Murray-Rust, P. Mining chemical information from open patents. *Journal of Cheminformatics* **3** (2011).
39. Lipinski, C. A. Drug-like properties and the causes of poor solubility and poor permeability. *Journal of Pharmacological and Toxicological Methods* **44**, 235–249 (2000).
40. Veber, D. F. *et al*. Molecular properties that influence the oral bioavailability of drug candidates. *Journal of Medicinal Chemistry* **45**, 2615–2623 (2002).
41. Lipinski, C. A., Lombardo, F., Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced Drug Delivery Reviews* **64**, 4–17 (2012).
42. Doak, B. C., Zheng, J., Dobritzsch, D. & Kihlberg, J. How beyond rule of 5 drugs and clinical candidates bind to their targets. *Journal of Medicinal Chemistry* **59**, 2312–2327 (2015).
43. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *Journal of Medicinal Chemistry* **39**, 2887–2893 (1996).
44. Baell, J. B. & Walters, M. A. Chemistry: Chemical con artists foil drug discovery. *Nature* **513**, 481–483 (2014).
45. Ermondi, G., Jiménez, D. G. & Sebastiano, M. R. Rational control of molecular properties is mandatory to exploit the potential of PROTACs as oral drugs. *ACS Medicinal Chemistry Letters* **12**, 1056–1060 (2021).
46. Jiménez, D. G. *et al*. Designing Soluble PROTACs: Strategies and preliminary guidelines. *Journal of Medicinal Chemistry* **65**, 12639–12649 (2022).
47. Ermondi, G., Jiménez, D. G. & Caron, G. PROTACs and building blocks: the 2D chemical space in very early drug discovery. *Molecules* **26**, 672 (2021).
48. Tashima, T. Proteolysis-Targeting Chimera (PROTAC) Delivery into the Brain across the Blood-Brain Barrier. *Antibodies* **12**, 43 (2023).
49. Xie, J. & Bogliotti, N. Synthesis and applications of Carbohydrate-Derived Macrocyclic Compounds. *Chemical Reviews* **114**, 7678–7739 (2014).
50. Zhao, Z. & Bourne, P. E. Rigid scaffolds are promising for designing macrocyclic kinase inhibitors. *ACS Pharmacology & Translational Science* **6**, 1182–1191 (2023).
51. Scott, K. A. *et al*. Stereochemical diversity as a source of discovery in chemical biology. *Current Research in Chemical Biology* **2**, 100028 (2022).
52. Chakravorty, S. J. *et al*. Nuisance compounds, PAINS filters, and dark chemical matter in the GSK HTS collection. *SLAS Discovery* **23**, 532–544 (2018).
53. Langdon, S. R., Brown, N. & Blagg, J. Scaffold diversity of exemplified medicinal chemistry space. *Journal of Chemical Information and Modeling* **51**, 2174–2185 (2011).
54. Malbon, J., Lawson, C. & Davison, M. The WTO Agreement on Trade-Related Aspects of Intellectual Property Rights: A Commentary. (Edward Elgar Publishing, 2014).
55. Motari, M. *et al*. 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** (2021).
56. Thakkar, K. & Billa, G. The concept of: Generic drugs and patented drugs vs. brand name drugs and non-proprietary (generic) name drugs. *Frontiers in Pharmacology* **4** (2013).
57. SCHULTZ, M. Pat-INFORMED: A new tool for drug procurement. *WIPO MAGAZINE* 30–36 (2018).
58. Senger, S., Bartek, L., Papadatos, G. & Gaulton, A. Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents. *Journal of Cheminformatics* **7** (2015).
59. Kim, S. *et al*. PubChem 2023 update. *Nucleic Acids Research* **51**, D1373–D1380 (2022).
60. Guo, Y. *et al*. An Integrated Strategy for Assessing the Metabolic Stability and Biotransformation of Macrocyclic Peptides in Drug Discovery toward Oral Delivery. *Analytical Chemistry* **94**, 2032–2041 (2022).
61. Münzfeld, L. *et al*. Synthesis and properties of cyclic sandwich compounds. *Nature* **620**, 92–96 (2023).
62. Gao, X. *et al*. Enantioselective Synthesis of Chiral Medium-Sized Cyclic Compounds via tandem Cycloaddition/Cope Rearrangement Strategy. *ACS Catalysis* **9**, 1645–1654 (2019).
63. Capuzzi, S. J., Muratov, E. & Tropsha, A. Phantom PAINS: Problems with the Utility of Alerts for Pan-Assay INterference CompoundS. *Journal of Chemical Information and Modeling* **57**, 417–427 (2017).

64. Senger, M. R., Fraga, C. A. M., Dantas, R. F. & Silva, F. P. Filtering promiscuous compounds in early drug discovery: is it a good idea? *Drug Discovery Today* **21**, 868–872 (2016).
65. Brown, D. G. An analysis of successful Hit-to-Clinical Candidate pairs. *Journal of Medicinal Chemistry* **66**, 7101–7139 (2023).
66. Cuschieri, S. Clinical trial publications. *Saudi Journal of Anaesthesia* **13**, 42 (2019).
67. Gadiya, Y., Zaliani, A., Gribbon, P. & Hofmann-Apitius, M. PEMT: a patent enrichment tool for drug discovery. *Bioinformatics* **39** (2022).
68. Gaulton, A. *et al*. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Research* **40**, D1100–D1107 (2011).
69. Wishart, D. S. *et al*. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Research* **46**, D1074–D1082 (2017).
70. Sydow, D., Morger, A., Driller, M. & Volkamer, A. TeachOpenCADD: a teaching platform for computer-aided drug design using open source packages and data. *Journal of Cheminformatics* **11** (2019).
71. Gadiya, Y. Dataset for manuscript titled "Exploring SureChEMBL from a drug discovery perspective". *Zenodo (CERN European Organization for Nuclear Research)* https://doi.org/10.5281/zenodo.10210061 (2023).

Acknowledgements

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. We would also like to thank Dr. Sarah Mubeen for helping us improve the language and content of the manuscript.

Author contributions

Y.G. conceived the work. Y.G. and A.Z. contributed to the ideation. Y.G. and S.S. performed the analysis. M.H.A., P.G., Y.G. and A.Z. have written the manuscript. All the authors have reviewed and approved the final manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Y.G.