Received: 30 August 2023 | Revised: 22 November 2023 | Accepted: 23 November 2023

DOI: 10.1002/minf.202300221

APPLICATION NOTE

molecular informatics

CIPSI: An open chemical intellectual property service for medicinal chemists

Maria Martinez-Sevillano[1] | **Maria J. Falaguera**[2,3] | **Jordi Mestres**[1,4]

[1]Systems Pharmacology, Research Group on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical Research Institute, Barcelona, Spain

[2]European Molecular Biology Laboratory – European Bioinformatics Institute (EMBL-EBI), Hinxton, UK

[3]Open Targets, Wellcome Genome Campus, Hinxton, UK

[4]Institut de Quimica Computacional i Catalisi, Facultat de Ciencies, Universitat de Girona, Girona, Spain

Correspondence
Jordi Mestres, Systems Pharmacology, Research Group on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical Research Institute, Doctor Aiguader 88, 08028 Barcelona, Spain.
Email: jmestres@imim.es and jordi.mestres@udg.edu

Funding information
Spanish Ministry of Science and Innovation, Grant/Award Number: PDC2021-120938-I00

Abstract

The availability of patent chemical data offers public access to a chemical space that is not well covered by other sources collecting small molecules from scholarly literature. However, open applications to facilitate the search and analysis of biologically-relevant molecular structures present in patents are still largely missing. We have developed CIPSI, an open Chemical Intellectual Property Service @ IMIM to assist medicinal chemists in searching and analysing molecules in SureChEMBL patents. The current version contains 6,240,500 molecules from 236,689 pharmacological patents, of which 5,949,214 are confidently assigned to core chemical structures reminiscent of the Markush structure in the patent claim. The platform includes some graphical tools to facilitate comparative patent analyses between drugs, chemical substructures, and company assignees. CIPSI is available at https://cipsi.org.

KEYWORDS
Patent chemistry, SureChEMBL, core chemical structures, patent annotation, comparative patent analyses

1 | INTRODUCTION

Patents represent important sources of biologically-relevant chemical series of molecules [1,2]. Nowadays, the major repository of patent chemical data is PubChem [3]. It contains over 78 million chemical structures from seven different patent data suppliers, of which around 17 million are exclusive patent compounds [4]. However, most chemical structures compiled from those suppliers are retrieved from patent texts and images using fully automated extraction methods [5]. This ensures a high volume of patent processing, but it also carries two important limitations. On one hand, there is no distinction between the compounds covered by patent claims and all sorts of starting materials, intermediate products, chemical reagents, and even solvent structures. A study on a subset of SureChEMBL pharmacological patents revealed that, on average, up to 35% of all molecules assigned to a patent were not relevant for the claim [6]. On the other hand, some of the chemical structures may contain errors (e.g., fragmentation, incompleteness, disconnection, wrong atom and/or bond types, open rings) due to failures in the automated extraction process. A recent analysis on a limited sample of PubChem patent compounds showed that errors in chemical structures automatically extracted from Patentscope (https://www.

wipo.int/patentscope/en/) and SureChEMBL [7] were observed in 60% of the cases [4]. These limitations contribute to the fact that patent chemical data is yet to be fully exploited in drug discovery.

One aspect that could facilitate access and usage of patent chemical data among the scientific community is the development of tools designed specifically to analyse them. As an example, a novel open-source patent enrichment tool was recently developed to assist in the extraction of relevant patent information linked to chemical structures and/or gene names described through FAIR principles and metadata annotations and by doing so support drug discovery to establish a patent landscape around genes of therapeutical interest [8,9]. This notwithstanding, open analytics tools to search and browse patent chemical data do not exist at present and they would greatly benefit the ability of drug discovery scientists to inspect visually all compounds and associated data in patents.

Here, we introduce CIPSI, an open Chemical Intellectual Property Service @ IMIM to analyse the chemical contents of SureChEMBL pharmacological patents [6]. The tool has been purposely designed for medicinal chemists and it includes the ability to search for active ingredients (names and InChIKeys), chemical substructures, company names and patent identifiers. We demonstrate its unique applicability to perform comparative patent analyses.

2 | MATERIALS AND METHODS

2.1 | Data collection

In this first release, CIPSI was built from the data contents of SureChEMBLccs v2021 (https://ftp.ebi.ac.uk/pub/databases/chembl/SureChEMBLccs/), a subset of the SureChEMBL public repository of US patent documents, including both patent applications and granted patents with priority date up to 2018 [7], enriched with molecules of pharmacological relevance. In total, it contains 6,240,500 molecules from 236,689 pharmacological patents (A61 K* IPC code), of which 5,949,214 are confidently assigned to core chemical structures and form highly congeneric chemical series [2,6]. The repository also includes 1,202,694 core chemical structures that represent the maximum common substructure identified among all molecules exemplified in 188,795 patent documents and that is taken as a substructural approximation to the patent claim [6]. A list of 251 company names has been assembled and mapped as assignees to 36,854 granted patents following a strict manual curation process (Supplementary Figure S1).

The data schema is composed of four main tables (Supplementary Figure S2), namely, molecules, core chemical structures, patents, and assignees. Each table contains a list of relevant attributes, most of which are extracted directly from the patent contents but others are computationally generated. Among the latter, molecule descriptors were obtained using the RDKit database cartridge v3.8 (https://www.rdkit.org/docs/Cartridge.html). All data were stored in a PostgreSQL v12.15 server.

2.2 | Backend architecture

The backend architecture of CIPSI consists of a Flask application v2.2.2 (https://flask.palletsprojects.com/en/2.2.x/) served with a Gunicorn v20.1.0 WSGI HTTP Server for UNIX (https://gunicorn.org/) built in a Python 3.7.7 environment (https://www.python.org/). The application also uses the RDKit library v2020.03.2.0 (https://www.rdkit.org/) to handle requests involving molecules. The psycopg2 library v2.9.3 (https://pypi.org/project/psycopg2/2.9.3/) is a python PostgreSQL adapter that allows the application to connect to the CIPSI database and run queries (Supplementary Figure S3).

2.3 | Frontend development

The frontend of the CIPSI application was developed using an Angular framework v13.3.9 (https://angular.io/), using components from Angular Material (https://material.angular.io/). Echarts graphical libraries were used to generate barplots (https://echarts.apache.org/) and word clouds (https://github.com/ecomfe/echarts-wordcloud). The MarvinJS applet from ChemAxon (https://chemaxon.com/marvin) is used as the molecular editor for (sub)structural searches. The frontend retrieves information from the backend via HTTP requests and responses (Supplementary Figure S3).

2.4 | Patent analytics

Patent dashboards have been designed to display all attributes in patent documents and associated chemical structure information, including the separation of molecules that contain the core chemical structure of the patent from the rest (Supplementary Figure S4). In addition, molecule, substructure, and assignee dashboards provide information on (i) the annual distribution of the priority dates from granted patents associated with the search, (ii) the concept landscape of the granted patent set, illustrated as a word cloud constructed based on a controlled

vocabulary of chemical, protein, side effect, and disease terms identified in the titles and abstracts of granted patents, and (iii) the complete list of patent documents (patent applications and granted patents), organised as a table containing the various patent attributes with both internal and external links to the patent dashboard and the corresponding PubChem page, respectively (Supplementary Figure S5). In all cases, a variety of filters can be applied to focus on a particular subset of patent documents.

3 | USE CASE: COMPARATIVE PATENT ANALYSES

Besides searching and browsing for patents, chemical (sub)structures, and companies, one of the unique features implemented in CIPSI is the ability to perform high-level comparative patent analyses between pairs of concepts within the same category. Four illustrative use cases are presented next.

3.1 | Comparing two patents

Patents US5466823A (Searle & Co.) and US5710140A (Merck & Co.) were granted to protect the chemical space around celecoxib and rofecoxib, respectively. CIPSI allows for comparing the distribution of up to six calculated molecular descriptors (molecular weight, hydrogen bond donors, hydrogen bon acceptors, octanol-water partition coefficient, number of rotatable bonds and polar surface area) from all molecules in two patents. In this case, one can observe that most of the 336 molecules in US5466823A have two hydrogen bond donors, a pharmacophoric feature absent in the majority of the 183 molecules in US5710140A (Supplementary Figure S6). The sulfonamide group present in the core chemical structure of the celecoxib analogues in patent US5466823A is most likely responsible for the difference encountered between the two hydrogen bond donor distributions.

3.2 | Comparing two drugs

Aripiprazole (ATC code N05AX12) and loratadine (ATC code R06AX13) are two marketed drugs used mainly for the treatment of schizophrenia and allergies, respectively. The comparative word cloud obtained in CIPSI from the set of patent documents containing these two drugs (Figure 1) clearly identifies their primary indications but also highlights the differences of the conceptual landscapes around them. The concept "schizophrenia" is detected in 6.5% of the 2788 aripiprazole patent documents (compared to only 0.4% of loratadine), whereas the concept "asthma" is found in 8.3% of the 4414 loratadine patent documents (in contrast to not being present in the word cloud of aripiprazole). These numbers can be easily obtained by passing the mouse

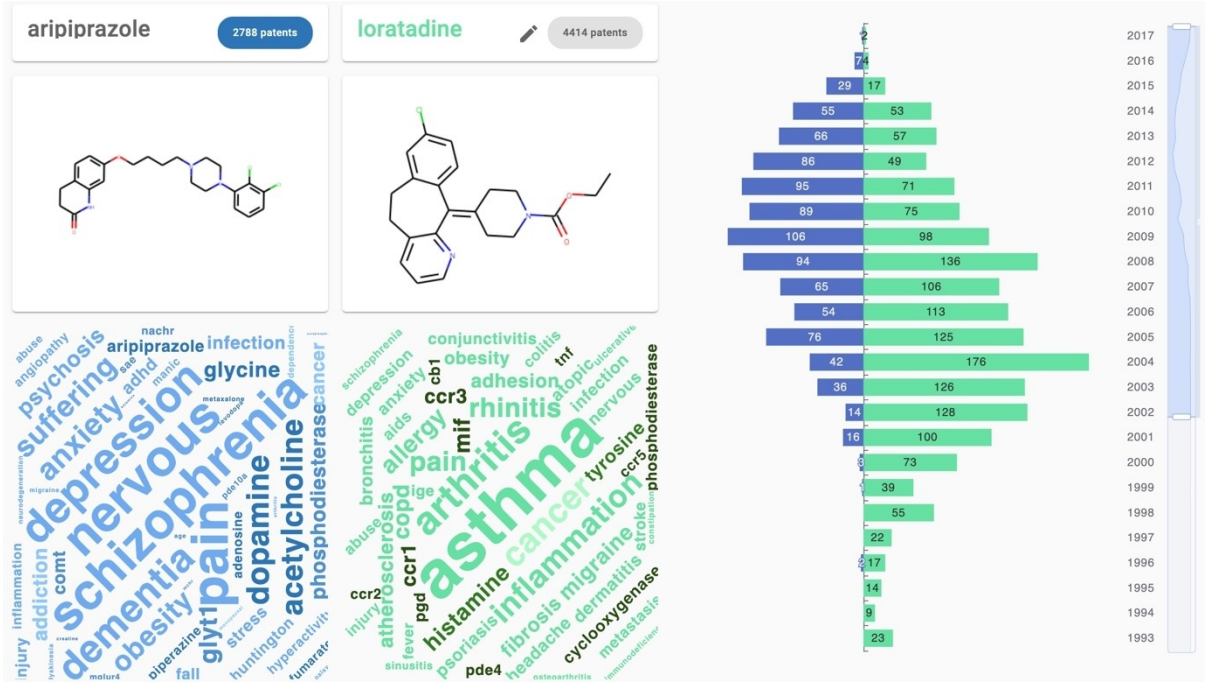

FIGURE 1 Comparative patent analysis between two drugs, aripiprazole and loratadine.

through the concepts on the word clouds. Note that the total patent counts, 2788 and 4414 patent documents containing aripiprazole and loratadine, respectively, consider both patent applications and granted patents. The word clouds and the bar chart annual distributions are however derived from the titles/abstracts and priority dates, respectively, of the corresponding subset of granted patents.

Conforming the "schizophrenia" landscape, the word cloud of aripiprazole reveals other related disease/safety concepts, such as "pain", "nervous", "dementia", "depression" and "anxiety", but also includes some molecule concepts as well, like "dopamine", "acetylcholine" and "glycine". In contrast, the "asthma" landscape of loratadine contains disease/safety concepts, such as "arthritis", "inflammation", "rhinitis" and "cancer", and molecule concepts, like "histamine" and "tyrosine", completely different from those found in the aripiprazole word cloud. However, when aripiprazole is compared to risperidone (ATC code N05AX08), another drug used in schizophrenia, the word clouds are remarkably similar (Supplementary Figure S7).

3.3 | Comparing two substructures

Chemical substructure searches in CIPSI return all patent documents that have at least one molecule containing the substructure. This patent set can then be compared to the corresponding patent set with molecules containing another substructure. As an illustrative example, we performed a comparative patent analysis between 1,3-diphenylurea and N,2-diphenylacetamide (Supplementary Figure S8). As can be observed, one simple atom change in the substructure alters the concept word cloud. The concept "pain" is found in 9.7% of the 694 patent documents that contain at least one compound with a 1,3-diphenylurea substructure, twice as much than its presence in 4.8% of the 588 patent documents containing compounds with a N,2-diphenylacetamide substructure. A closer look at the word cloud identifies concepts like "infection", that show similar patent frequencies (3.5% in 1,3-diphenylurea *versus* 3.2% in N,2-diphenylacetamide), whereas other concepts like "nervous" and "obesity" appear to be exclusive of patent documents containing compounds with 1,3-diphenylurea and N,2-diphenylacetamide, respectively.

3.4 | Comparing two companies

A comparative analysis of the concept landscapes extracted from the title and abstracts of SureChEMBL patent documents assigned to AstraZeneca and Novartis was performed (Supplementary Figure S9). As can be observed, "cancer" is the most salient concept appearing in both word clouds. It is found in 11.9% of the 1370 patents from AstraZeneca and 11.1% of the 1177 patents from Novartis. Other concepts like "arthritis" are also equally encountered in patents from both companies (3.7% in AstraZeneca *versus* 4.2% in Novartis). However, other concepts are over-represented in one of the companies or even company exclusive. For example, disease/safety concepts, such as "pain", "dementia", "thrombosis", "cholesterol" and "incontinence", and molecule concepts, like "omeprazole", are over-represented or exclusive of AstraZeneca, whereas disease/safety concepts, such as "hypertension", "diabetic", "fibrosis", "inflammation", "osteoporosis", and "retinopathy", and molecule concepts, like "valsartan" and "imatinib", are over-represented or exclusive of Novartis.

4 | DISCUSSION AND FUTURE WORK

CIPSI is an open chemical intellectual property service purposely designed to assist medicinal chemists in the search for drugs and chemical substructures in SureChEMBL pharmacological US patents. The potential applicability of CIPSI ranges from enabling manual curation efforts of patent chemical data to acting as a source of competitive intelligence. In this sense, it is unique in its ability to perform visual comparative patent analyses between molecular entities.

However, the current version is not exempt of limitations. For example, some concepts in word clouds may appear overweighted by the existence of multiple US patent documents belonging to the same patent family and some bar charts may be slightly distorted by accounting for multiple US granted patents that in fact belong to the same patent family. Plans for a future version contemplate addressing these limitations and extending patent coverage from SureChEMBL to other available patent sources, such as Espacenet (https://worldwide.espacenet.com/).

ACKNOWLEDGMENTS

This work was supported by the Spanish Ministry of Science and Innovation under the "Proof of Concept" project grant PDC2021-120938-I00.

CONFLICT OF INTEREST STATEMENT

JM is the co-founder and co-owner of the company Chemotargets. The other authors declare no competing interests.

DATA AVAILABILITY STATEMENT

The patent data that support the findings of this study are openly available at https://ftp.ebi.ac.uk/pub/databases/chembl/SureChEMBLccs/. The CIPSI platform is accessible at http://cipsi.imim.es.

ORCID

Jordi Mestres http://orcid.org/0000-0002-5202-4501

REFERENCES

1. C. Southan, *Drug Discov. Today Technol.* **2015**, *14*, 3–9.
2. M. J. Falaguera, J. Mestres, *Molecules* **2021**, *26*, 5253.
3. S. Kim, *et al., Nucleic Acids Res.* **2023**, *51*, D1373–D1380.
4. J. Ohms, *World Patent Inf.* **2022**, *70*, 102134.
5. J. Ohms, *World Patent Inf.* **2021**, *66*, 102055.
6. M. J. Falaguera, J. Mestres, *J. Chem. Inf. Model.* **2021**, *61*, 2241–2247.
7. G. Papadatos, *et al., Nucleic Acids Res.* **2016**, *44*, D1220–D1228.
8. Y. Gadiya, *et al., Bioinformatics* **2023**, *39*, btac716.
9. Y. Gadiya, *et al., Artif. Intel. Life Sci.* **2023**, *3*, 100069.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: M. Martinez-Sevillano, M. J. Falaguera, J. Mestres, *Molecular Informatics* **2024**, *43*, e202300221. https://doi.org/10.1002/minf.202300221