

Identification of the Core Chemical Structure in SureChEMBL Patents

Maria J. Falaguera and Jordi Mestres*

Research Group on Systems Pharmacology, Research Program on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical Research Institute and University Pompeu Fabra, Parc de Recerca Biomèdica (PRBB), Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain.

ABSTRACT

The SureChEMBL database provides open access to 17 million chemical entities mentioned in 14 million patents published since 1970. However, alongside with molecules covered by patent claims, the database is full of starting materials and intermediate products of little pharmacological relevance. Herein, we introduce a new filtering protocol to automatically select the core chemical structures best representing a congeneric series of pharmacologically relevant molecules in patents. The protocol is first validated against a selection of 890 SureChEMBL patents for which a total of 51,738 manually curated molecules are deposited in ChEMBL. Our protocol was able to select 92.5% of the molecules in ChEMBL from all 270,968 molecules in SureChEMBL for those patents. Subsequently, the protocol was applied to all 240,988 US pharmacological patents for which 9,111,706 molecules are available in SureChEMBL. The unsupervised filtering process selected 5,949,214 molecules (65.3% of the total number of molecules) that form highly congeneric chemical series in 188,795 of those patents (78.3% of the total number of patents). A SureChEMBL version enriched with molecules of pharmacological relevance is available for download at <ftp://ftp.ebi.ac.uk/pub/databases/chembl/SureChEMBLccs>.

* To whom correspondence should be addressed. Email: jmestres@imim.cat

1 Introduction

Pharmacological patents are a key source of information in drug discovery as they offer early access to novel chemical space of biological relevance. Motivated by the competitiveness of the business sector, the patent system encourages the constant discovery and disclosure of new active structures,¹ often poorly covered in the scientific literature.² In this respect, a recent comparison between patent-derived and literature-extracted data revealed that, from the 15.4 million chemical structures available in all large patent-derived chemical sources, only 0.5 million were found to be present also in literature-derived databases.³ And when comparing the deposition date in patents of these 0.5 million molecules with their corresponding publication date in scholar literature, an average lag time of four years was observed,² with delays going up to six years for its final storage in publicly available sources such as ChEMBL.⁴ Therefore, there is a need for an early, more complete and accurate open access to molecules exemplified in pharmacological patents.

For years, access to chemical information published in patents was only possible through commercial databases such as CAS SciFinder, Excelra GOSTAR, Elsevier Reaxys, or Thomson Reuters Pharma,¹ which guarantee manually curated, regularly updated data.³ Alternatively, other sources such as SCRIpDB,⁵ IBM contribution to the US National Institutes of Health (NIH),⁶ ChEMBL⁴ and PubChem⁷ offer open access to patent chemical data of pharmacological relevance, although the first two have not been updated for years and the patent coverage is generally limited compared to their commercial counterparts.¹

But in 2016, open access to patent chemical data changed completely with the publication of SureChEMBL,¹ a database derived from SureChem,⁸ a commercial product with significantly wider patent coverage than most of the other patent chemical databases. In its first release (April 2016), SureChEMBL contained 17 million chemical structures from 14 million patents published since 1970 from all three major patent authorities, namely, the World Intellectual Property Organization (WIPO), the United States Patent and Trademark Office (USPTO), and the European Patent Office (EPO). Apart from chemical structures, SureChEMBL provides patent titles, International Patent Classification (IPC) codes (IPCPUB v8.0, WIPO) and it is regularly updated.¹

The high patent coverage of SureChEMBL compared to other chemical databases of its kind is the result of applying automated chemical named entity recognition technology to extract every chemical structure from text, images and MOL files attached to the patent document.³ This process ensures the identification and extraction of all chemical entities mentioned in patents. However, this is also one of the recognized limitations of SureChEMBL, as there is no distinction between starting materials, intermediate products and pharmacologically relevant compounds, all ultimately being deposited in the database. To address this situation, Kunimoto and Bajorath⁹ applied the matched molecular pairs (MMP) concept to detect the main substructure shared by the small molecules contained in a patent claim. More recently, Akhondi *et al.*¹⁰ developed a text-mining recognition system for relevant compounds in a patent based on analyzing the patent's context of a compound defined by its position in the document, the section where it appears, the frequency of appearance, its wide usage in other patents, and any other compounds being mentioned in its textual vicinity. In spite of these efforts, a fully automatic and efficient process to detect molecules of therapeutic relevance in SureChEMBL patents is still missing.

Here we introduce a new filtering protocol to identify the core chemical structure in SureChEMBL patents and extract all pharmacologically relevant molecules exemplifying the patent claims. The approach is validated on its ability to automatically extract the manually curated subset of compounds from 890 SureChEMBL patents present in ChEMBL. Subsequently, the protocol is applied to all 240,988 pharmacological patents from the United States (US) covered in SureChEMBL. The final subset of filtered SureChEMBL molecules from US pharmacological patents is available at the EMBL-EBI website.¹¹

2 Methods

2.1 SureChEMBL database

In the release used in this work (July, 2019), SureChEMBL covered 1,975,722 US patents containing 167,662,929 patent-molecule associations involving 14,284,051 unique small molecules. Out of this total number of US patents, 240,988 (12.2%) can be considered “pharmacological” patents, which contain 45,539,938 patent-molecule associations with 9,111,706 unique small molecules. We define a patent as “pharmacological” when it has an A61K* IPC code, with the exception of A61K6 (preparations for dentistry), A61K7 or A61K8 (cosmetics or similar toilet preparations), A61K9 (medicinal preparations characterised by special physical form), A61K38 (medicinal preparations containing peptides), A61K39 (medicinal preparations containing antigens or antibodies) and A61K48 (medicinal preparations containing genetic material which is inserted into cells of the living body to treat genetic diseases). However, patents tend to have multiple IPC codes to describe their uses and applications. As illustrated in Fig. 1a, the most frequent classification code in US A61K* patents is A61K31 (medicinal preparations containing organic active ingredients), but annotations to non-A61K* codes, such as A61P25 (drugs for disorders of the nervous system), C07D401 (heterocyclic compounds containing two or more hetero rings, having nitrogen atoms as the only ring hetero atoms), and C07D413 (heterocyclic compounds containing two or more hetero rings, at least one ring having nitrogen and oxygen atoms as the only ring hetero atoms), are also frequently encountered.

The most repeated terms present in the title of pharmacological patents are shown in Fig. 1b. Among those, words such as “derivatives”, “inhibitor”, “compounds”, “active”, or “modulator” reflect the main underlying nature of the compounds claimed by those patents. But patent compounds collected in SureChEMBL do not include only claimed bioactive small molecules but also common reactants, intermediate products, inorganic compounds, and any other small molecules mentioned in patent files. In this respect, SureChEMBL patents have over one order of magnitude (x12) more patent-molecule associations than unique small molecules, clearly reflecting the existence of some molecules frequently included in multiple patents. Interestingly, this promiscuity is significantly reduced (x5) in pharmacological patents.

a)



b)



Fig. 1. Word clouds showing the most frequent **a)** IPC codes and **b)** title terms in SureChEMBL pharmacological patents. The larger the word font, the more frequent the word is in patents.

2.2 Filtering protocol

A filtering protocol was implemented to identify the set of pharmacologically relevant molecules covered by the patent claim from all molecules of a given patent. The protocol is based on the assumption that all relevant compounds in a patent share a core chemical structure that may be represented by an ensemble of candidate maximum common substructures (MCSs) and that these candidate MCSs are significantly more populated with similar congeneric compounds than any other MCS identified from the other compounds in the patent. The entire process includes three filtering steps and two additional refinement steps.

(1) Extraction of MCSs. Using as input the SMILES of all compounds in a SureChEMBL patent, the first step is to extract the MCSs for all pairwise combinations of compounds. For this, the `rdkit.Chem.rdmcs.FindMCS` function is used with the parameters `RingMatchesRingOnly` and `CompleteRingsOnly` activated. A total of 10,377,468 unique MCSs were extracted from all 240,988 US pharmacological patents. At this stage, a promiscuity value, defined as the number of patents in which a given MCS is found, is also assigned to each MCS. About 59% of all unique MCSs are found exclusively in a single patent, whereas less than 3% are present in 10 or more patents.

(2) Deletion of promiscuous MCSs. The main objective of this second step is to discard all molecules in patents likely to be associated with reactants and other substances commonly used in chemistry and thus, unrelated to the patent claims. To this aim, all molecules containing MCSs found above the 1-quantile of the distribution of associated patent promiscuities within a patent were discarded. About 46% of the patents retained only molecules with MCSs exclusive to them. In contrast, almost 33% of the patents admitted molecules containing MCSs with promiscuities ranging from 1 to 10 or higher.

(3) Selection of candidate MCSs. This third step aims at identifying the ensemble of MCSs that are most likely to represent the core chemical structure of the patent claim. Three properties of the molecules defining each MCS are considered: i) coverage, calculated as the percentage of patent molecules containing the MCS; ii) homogeneity, calculated as the average pairwise Tanimoto similarity between the RDKit fingerprints of all molecules sharing a MCS; and iii) inclusion, measured as the percentage of all other MCSs found to be substructures of a given MCS. Then, a final score reflecting the properties of the chemical space of each MCS (MCScore) is calculated as $\text{MCScore} = \text{coverage} * \text{homogeneity} * \text{inclusion}$. Once scored, for a MCS to be considered as a candidate MCS likely to reflect the core chemical structure of the patent claim, its MCScore needs to be equal or greater than the 70-quantile threshold of the distribution of MCScores in the patent. At the end of this step, from all molecules of a pharmacological patent in SureChEMBL, only those molecules associated with at least one of the candidate MCSs will be retained for further consideration.

(4) Recovery of highly similar molecules. Singular molecules representing some low coverage and slightly heterogenous MCS, yet highly similar to molecules from those candidate MCSs selected in the previous step, can still be recovered here if the pairwise Dice similarity between their Morgan fingerprints and those of any of the previously selected molecules is equal or greater than 80%.

(5) Selection of high confidence patents. This fifth step is added to assign a confidence label to each patent based on the degree of structural congenericity of the final selected molecules. In this respect, under the assumption that molecules exemplifying a patent claim should define a close congeneric chemical series, the median value of the distribution of pairwise Dice similarities between all patent molecules will be associated with the level of confidence on the patent. Based on a validation analysis (*vide infra*), patents will be given a “high confidence” flag if the median similarity value is equal or higher than 40%.

3 Results and Discussion

3.1 Validation against SureChEMBL patents in ChEMBL

In order to validate the performance of the filtering protocol on its ability to extract claimed molecules from SureChEMBL patents, we took the highly curated set of molecules available in ChEMBL_23 (May, 2017) extracted from a selected number of those patents. We found a total of 51,738 molecules annotated with *in vitro* pharmacology data in ChEMBL coming from 890 SureChEMBL US A61K* pharmacological patents. However, there are 270,968 molecules in SureChEMBL associated with those same 890 patents. Therefore, the challenge is to assess to which extent an unsupervised filtering protocol is able to automatically retrieve those 51,738 molecules from the pool of 270,968 molecules. The results are compiled in Table 1.

Table 1. Filtering protocol applied to the 890 SureChEMBL US pharmacological patents included in ChEMBL. The number of (and percentage from total) patents, molecules in ChEMBL and corresponding molecules in SureChEMBL left at each filtering step is provided.

Filtering step	No. patents (% from total)	No. molecules in ChEMBL (% from total)	No. molecules in SureChEMBL (% from total)
(0) SureChEMBL@ChEMBL	890 (100.0%)	51,738 (100.0%)	270,968 (100.0%)
(1) Extraction of MCSs	889 (99.8%)	51,737 (100.0%)	270,967 (100.0%)
(2) Deletion of promiscuous MCSs	889 (99.8%)	49,464 (95.6%)	192,492 (71.0%)
(3) Selection of candidate MCSs	889 (99.8%)	43,931 (84.9%)	142,414 (52.6%)
(4) Recovery of highly similar molecules	889 (99.8%)	48,335 (93.4%)	163,091 (60.2%)
(5) Selection of high confidence patents	851 (95.6%)	47,857 (92.5%)	159,439 (58.8%)

As it can be observed, MCSs can be extracted from all patents except one. This is patent US-8685986 claiming a medical composition for treatment or prophylaxis of glaucoma and for which only one molecule was extracted from its abstract, namely, 2-(pyridine-2-ylamino)acetic acid. Since you need at least a pair of molecules to define a MCS, that patent was dropped at the very first step.

The second step, involving the removal of molecules associated with promiscuous MCSs, is the one having the strongest filtering effect. A total of 78,475 molecules are discarded, which correspond to 61.0% of all molecules that will be ultimately filtered out. At this stage, we are left with 192,492 molecules, 71.0% of the initial number of SureChEMBL molecules, which nonetheless include 49,464 molecules present in ChEMBL, that is, 95.6% of all molecules in ChEMBL for those patents.

Selecting molecules from candidate MCSs only results also in an important reduction of the number of molecules kept from patents. A total of 50,078 molecules are excluded in this third step, which correspond to 39.0% of all molecules discarded at the end of the three filtering steps. The number of molecules remaining at this stage is 142,414, almost half (52.6%) of the initial number of SureChEMBL molecules. Within them, there are 43,931 molecules present in ChEMBL, which represent 84.9% of all molecules in ChEMBL for those patents.

Applying similarity criteria to identify molecules that may have been discarded at any of the previous steps because of their relatively high MCS promiscuity or low coverage, homogeneity and inclusion values of their MCSs results in the recovery of 20,677 molecules. This increases the number of molecules retained at this fourth step up to 163,091, 60.2% of the initial number of SureChEMBL molecules, which include 48,335 molecules present in ChEMBL, 93.4% of all molecules in ChEMBL for those 890 patents.

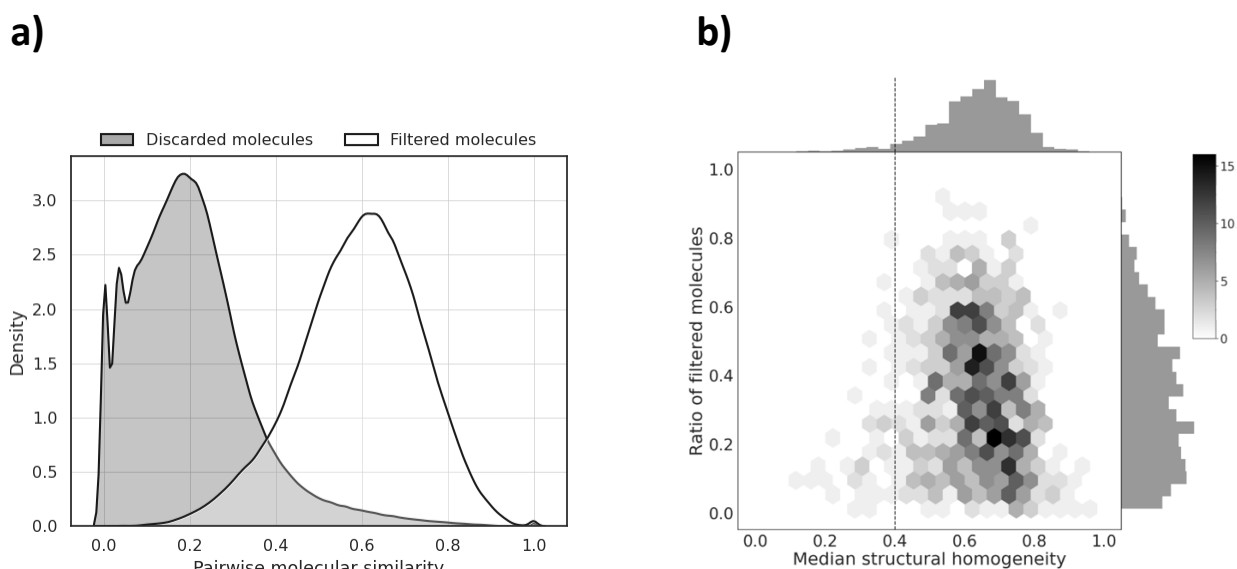


Fig. 2. a) Kernel density plot for the distribution of pairwise similarities between the 163,091 filtered molecules (white surface) and the 107,877 discarded molecules (grey surface) up to step 4 of the filtering protocol; **b)** Density plot of median structural homogeneity values against the ratio of filtered molecules in patents. Grey scale of hexagons corresponds to the relative density of patents. Also included are the distributions of the number of patents corresponding to each median structural homogeneity (top x-axis) and each ratio of filtered molecules (right y-axis). The dotted line at a median structural homogeneity of 0.4 marks the threshold for high confidence patents.

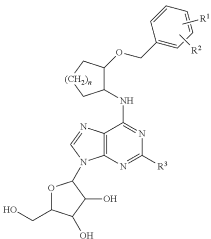
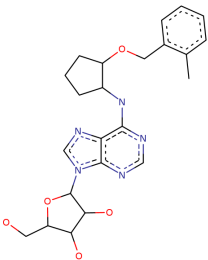
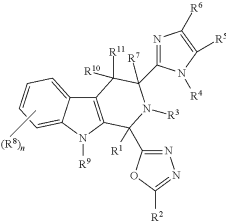
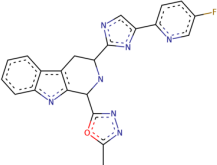
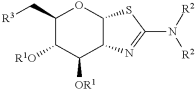
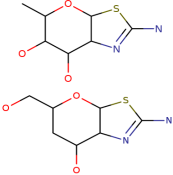
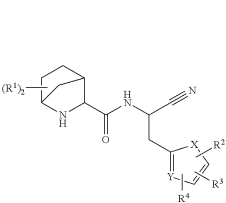
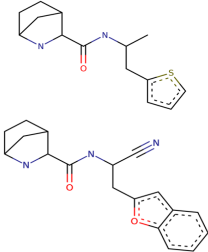
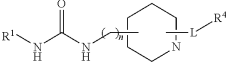
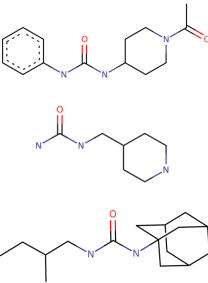
The 163,091 molecules that passed all filters (filtered molecules) within each patent up to this stage should have clearly a higher degree of congenericity than the 107,877 molecules that did not pass any of the filters (discarded molecules). To confirm this assumption, kernel density plots of the pairwise similarity distributions for filtered and discarded molecules were compared (Fig. 2a). As it can be observed, there is a clear separation between the two sets, with similarity values at the density peaks of the distributions being 0.60 and 0.19 for filtered and discarded molecules, respectively. A more in-depth analysis would involve adding another dimension to reflect the ratio of filtered molecules remaining in the end within each patent (Fig. 2b). As it is shown, most patents have median structural homogeneities between 0.5 and 0.8 and retain between 10% and 60% of the original molecules in SureChEMBL.

A close look at patents having median structural homogeneity values below 0.4 (Fig. 2b) revealed that their filtered molecules come from multiple candidate MCSs that may define different regions of a large Markush structure or simply different congeneric series. For these patents, visual inspection of their filtered molecules would be strongly advised. Accordingly, a homogeneity value of 0.4 was established as the lower-bound threshold to identify patents with a high degree of confidence that the final filtered molecules reflect a congeneric series of a well-defined and consistent patent claim. When this threshold was implemented as the last step of the filtering protocol (Table 1), a total of 38 patents were affected, leaving a final number of 851 high-confidence patents, 95.6% of the initial SureChEMBL patents in ChEMBL. This affected 3,652 molecules in SureChEMBL, leaving the final count of filtered molecules to 159,439, 58.8% of the initial number of SureChEMBL molecules for those patents. This means that over 40% of molecules in those SureChEMBL patents are likely common reactants, intermediate products, and other

small molecules not relevant for the patent claims. In contrast, the filtering protocol was able to retain 47,857 molecules present in ChEMBL, that is, 92.5% of all pharmacologically relevant molecules carefully selected and included in ChEMBL from those SureChEMBL patents.

Results for some illustrative examples of high-confidence patents are collected in Table 2. One of them is patent US-8501708 that aims at protecting a class of purine nucleoside compounds as selective A1 adenosine receptor agonists. A total of 115 molecules are present in SureChEMBL. For this particular patent, we found a perfect match between the Markush structure provided in the claim and the only candidate MCS contained in 69 molecules (60% of total) that form a highly congeneric series (median similarity of 0.87). Among them, all 18 molecules (16% of total) contained in ChEMBL were recovered. Another example is patent US-8754099 that protects beta-carboline derivatives as selective antagonists of the somatostatin subtype receptor 3 for the treatment of type-2 diabetes. In this case, SureChEMBL contains 184 molecules, of which 26 (14%) were found to define a highly congeneric series (median similarity of 0.80) around a candidate MCS that matches nicely the Markush structure of the claim. All 4 molecules (2% of total) present in ChEMBL were found within the 26 molecules selected by the filtering protocol. A third example is provided by patent US-8933040 that protects a series of compounds as selective glycosidase inhibitors. Of the 170 molecules in SureChEMBL, 57 molecules (33% of total) were selected by the filtering protocol, among which all 6 molecules (3% of total) in ChEMBL were present. Note that in this case, not just one but two candidate MCS were identified, both of which match well with the Markush structure of the claim. Another example of a two-candidate MCS case is patent US8871783 that protects the use of 2-aza-bicyclo[2.2.1]heptane-3-carboxylic acid (cyano-methyl)-amides as cathepsin C inhibitors. A total of 255 molecules were found in SureChEMBL for this patent. Of them, 58 molecules (22% of total) passed all steps of the filtering protocol among which all 26 molecules (10%) present in ChEMBL were recovered. Finally, an example of a case for which the filtering protocol selected three candidate MCS is offered by patent US-8501783. Interestingly, the abstract of the patent states that the invention relates to inhibitors of the soluble epoxide hydrolase that incorporate multiple pharmacophores, therefore justifying the need for multiple candidate MCS to identify all pharmacologically relevant molecules covered by the patent. Of the 190 molecules present in SureChEMBL, 66 molecules (35% of total) passed all filters. In this case, of the 72 molecules contained in ChEMBL for this patent, 51 molecules (71%) were contained within the 66 molecules selected. In general, for any given patent, beyond recovering most of the molecules already in ChEMBL, additional molecules belonging to the same chemical series were retrieved. This exemplifies the potential of the approach to produce a version of SureChEMBL containing only molecules around the main core chemical structures (ccs) identified in patents (SureChEMBLccs).

Table 2. Illustrative examples of SureChEMBL patents present in ChEMBL. The Markush structure is the one provided in the patent document. Performance of the filtering protocol is assessed in terms of ability to identify the core chemical structure and extract exemplified molecules covered in ChEMBL.

Patent ID (median similarity)	Markush structure in patent document	Candidate MCSs selected	Total no. SureChEMBL molecules	Selected no. SureChEMBL molecules (%total)	No. SureChEMBL molecules in ChEMBL (%total sensitivity)
US-8501708 (0.87)			115	69 (60%)	18 (16% 100%)
US-8754099 (0.80)			184	26 (14%)	4 (2% 100%)
US-8933040 (0.65)			170	57 (33%)	6 (3% 100%)
US-8871783 (0.64)			255	58 (22%)	26 (10% 100%)
US-8501783 (0.68)			190	66 (35%)	72 (38% 71%)

3.2 Application to all SureChEMBL US A61K patents

The 890 SureChEMBL patents covered in ChEMBL represent only 0.4% of all US A61K pharmacological patents in SureChEMBL. Having validated the performance of the filtering protocol on those 890 patents, the next step was to apply it to the entire set of 240,988 SureChEMBL patents. Since the filtering protocol takes on average 7 seconds per patent, such a large-scale application required extensive computational resources. In a cluster composed of 20 nodes, each having 96 Gb of RAM, 2 CPUs AMD Opteron™ Processor 6234, providing 24 cores, a GlusterFS distributed file system with 90 Tb of storage and using Slurm Workload Manager as queue batch system, all those SureChEMBL patents were processed in about 5 days. The results obtained are collected in Table 3.

The first filtering step of the protocol, involving the extraction of MCSs from molecules in the patent, resulted in a drop of 4,299 patents. There are essentially two main reasons why no MCS could be extracted for 1.8% of the patents. One of them is that, in some cases, there is a limited number of molecules extracted from the patent and these molecules are highly diverse. This is often due to the absence of formulas and images of the claimed molecules in the patent or due to the low-quality of the documents describing the patents. Indeed, patent documents prior to 2007 can contain low-quality chemical names and images that may hinder SureChEMBL's image and text mining procedures. In this respect, it is important that patent offices encourage applicants to submit chemical structure files of claimed molecules attached to the patent application document. On the other hand, some patents contain very large molecules that make MCS extraction extremely time consuming. To skip these cases, a time limit was imposed when attempting to extract MCSs from a given patent. Overall, a total of 938,170 molecules associated with these 4,299 patents were discarded, 27.8% of all molecules that will be filtered out along the process, leaving 8,173,536 molecules at this stage, 89.7% of the initial number of molecules.

The removal of molecules associated with promiscuous MCSs in a second step affected 768,406 molecules, 22.8% of all molecules removed by the filtering protocol. Molecules containing these promiscuous MCSs come from three main sources. The first group is composed of a heterogeneous set of molecules considered common reactants (e.g. EDTA, halazone, nitrophenyl phosphate and HEPES), inorganic compounds (such as polyalcohols), substituent groups (e.g. thiol, methyl and butane) and amino acids that are commonly present in patents claiming some heterogeneous formulations used as topical remedies, solutions containing an active principle or dialysis solutions, among others. The second group includes monosaccharides, nucleotides and its derivatives present frequently in patents claiming oligonucleotides for gene therapy, antibodies or other biologics. Molecules of this sort were not expected to be encountered, since we removed *a priori* all patents with IPC codes A61K39 (antibodies), A61K48 (genetic material) and A61K9 (physical forms). However, it was found that in some cases, especially for old patents, these classification codes were not as specific as expected. Finally, the third group contains a short list of very popular bioactive molecules that are included, either themselves or some derivatives of them, in the claim of patents for different uses, as ingredients of *in vivo* cell cultures and medical formulations, or appear as example drugs in the section that describes the background of the invention. Examples of such drugs are porphyrin (used as chelant in photodynamic therapy), fluorescein (used as diagnostic tool in the field of ophthalmology and optometry), staurosporine (used in cancer treatment), omeprazole and pantoprazole (used for stomach ulcer treatment), and vitamins (such as folic acid derivatives and ergocalciferol). A total of 7,405,130 molecules remained at this stage, 81.3% of the initial number of molecules.

Table 3. Filtering protocol applied to all 240,988 SureChEMBL US pharmacological patents. The number of (and percentage from total) patents and molecules in SureChEMBL left at each filtering step is provided.

Filtering step	No. patents (% from total)	No. molecules in SureChEMBL (% from total)
(0) SureChEMBL@ChEMBL	240,988 (100.0%)	9,111,706 (100.0%)
(1) Extraction of MCSs	236,689 (98.2%)	8,173,536 (89.7%)
(2) Deletion of promiscuous MCSs	236,689 (98.2%)	7,405,130 (81.3%)
(3) Selection of candidate MCSs	236,689 (98.2%)	5,736,478 (63.0%)
(4) Recovery of highly similar molecules	236,689 (98.2%)	6,240,500 (68.5%)
(5) Selection of high confidence patents	188,795 (78.3%)	5,949,214 (65.3%)

Retaining molecules from candidate MCSs only had the strongest filtering effect, with 49.4% (1,668,652 molecules) of all molecules discarded being removed in this third step. The number of molecules remaining at this stage is 5,736,478, 63.0% of the initial number of molecules. Subsequently, applying the similarity criteria defined above to identify molecules that may have been discarded previously because of the relatively high promiscuity or low coverage, homogeneity and inclusion values of their MCSs recovers 504,022 molecules. This increases the number of molecules retained up to 6,240,500 molecules, 68.5% of the initial molecules.

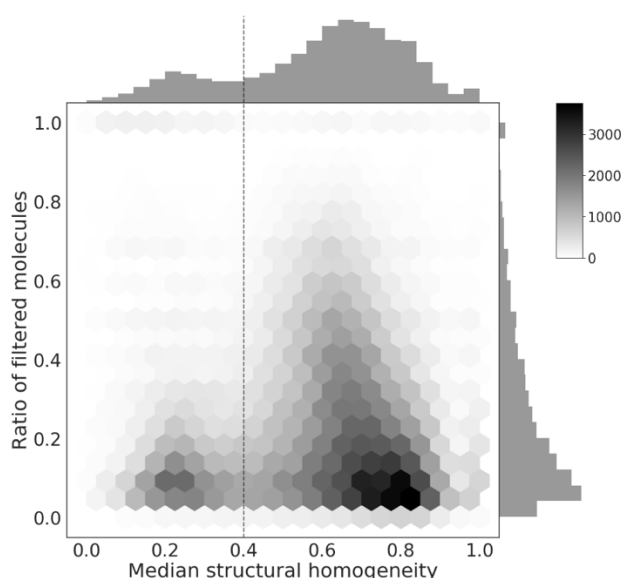


Fig. 3. Density plot of median structural homogeneity values against the ratio of filtered molecules in all US pharmacological patents in SureChEMBL. Grey scale of hexagons corresponds to the relative density of patents. Also included are the distributions of the number of patents corresponding to each median structural homogeneity (top x-axis) and each ratio of filtered molecules (right y-axis). The dotted line at a median structural homogeneity of 0.4 marks the threshold for high confidence patents.

Finally, based on the previous validation exercise, a median structural homogeneity threshold of 0.4 was applied to select the list of high confidence patents that contain sets of highly congeneric compounds (Fig. 3). The application of this filter affected 47,894 patents, 19.9% of the initial SureChEMBL patents, resulting in the removal of 291,286 molecules. In the end, a total of 5,949,214 molecules were left, 65.3% of all molecules from SureChEMBL US pharmacological patents considered originally.

4 Conclusion

With the advent of a new generation of artificial intelligence algorithms to recognize and extract chemical structures from patent documents in a more reliable and efficient manner,¹² unsupervised processes to confidently identify the subset of molecules covered by patent claims from all extracted chemical structures are required. In this work, a filtering protocol was designed to automatically select the core chemical structures best representing a congeneric series of pharmacologically relevant molecules in a patent. To demonstrate the validity of the approach, we applied it first to a set of 270,968 chemical structures from a selection of 890 SureChEMBL patents for which a total of 51,738 manually curated molecules are deposited in ChEMBL. Our protocol was able to identify and discard 41.2% of all molecules in SureChEMBL and retain, within the remaining 58.8%, up to 92.5% of all molecules in ChEMBL. In a second step, we performed a large-scale experiment against 240,988 US pharmacological patents for which 9,111,706 molecules are available in SureChEMBL. With a computational cost of approximately 5 days, our protocol selected 5,949,214 molecules (65.3% of the total number of molecules) that form highly congeneric chemical series in 188,795 of those patents (78.3% of the total number of patents). We believe that this protocol will be useful to assist in the process of producing regular updates of a SureChEMBL version enriched with molecules of pharmacological relevance for the benefit of the entire scientific community.

Availability

The SureChEMBL subset of molecules claimed by pharmacological US patents is available for download at <ftp://ftp.ebi.ac.uk/pub/databases/chembl/SureChEMBLccs>.

Conflicts of interest

There are no conflicts to declare.

References

1. G. Papadatos, M. Davies, N. Dedman, J. Chambers, A. Gaulton, J. Siddle, R. Koks, S. A. Irvine, J. Pettersson, N. Goncharoff, A. Hersey and J. P. Overington, *Nucleic Acids Res.*, 2016, **44**, D1220–D1228.
2. S. Senger, *J. Cheminform.*, 2017, **9**, 26.
3. C. Southan, *Drug Discov. Today Technol.*, 2015, **14**, 3-9.
4. D. Mendez, A. Gaulton, A. P. Bento, J. Chambers, M. De Veij, E. Félix, M. P. Magariños, J. F. Mosquera, P. Mutowo, M. Nowotka, M. Gordillo-Marañón, F. Hunter, L. Junco, G. Mugumbate, M. Rodriguez-Lopez, F. Atkinson, N. Bosc, C. J. Radoux, A. Segura-Cabrera, A. Hersey and A. R. Leach, *Nucleic Acids Res.*, 2019, **47**, D930–D940.
5. A. Heifets and I. Jurisica, *Nucleic Acids Res.*, 2012, **40**, D428–D433.
6. IBM Press Release, 2011. <http://www-03.ibm.com/press/us/en/pressrelease/36180.wss> (last accessed on January 28th, 2021).
7. Y. Wang, S. H. Bryant, T. Cheng, J. Wang, A. Gindulyte, B. A. Shoemaker, P. A. Thiessen, S. He and J. Zhang, *Nucleic Acids Res.*, 2017, **45**, D955–D963.
8. Digital Science Ltd. News Blog, 2013. <https://www.digital-science.com/blog/tag/surechem> (last accessed on January 28th, 2021).
9. R. Kunitomo and J. Bajorath, *J. Comput. Aided Mol. Des.*, 2017, **31**, 779–788.
10. S. A. Akhondi, H. Rey, M. Schwörer, M. Maier, J. Toomey, H. Nau, G. Ilchmann, M. Sheehan, M. Irmer, C. Bobach, M. Doornenbal, M. Gregory and J. A. Kors. *Database (Oxford)*, 2019, baz001.
11. SureChEMBLccs, 2021. <ftp://ftp.ebi.ac.uk/pub/databases/chembl/SureChEMBLccs> (last accessed on January 28th, 2021).
12. J. Staker, K. Marshall, R. Abel and C. M. McQuaw. *J. Chem. Inf. Model.*, 2019, **59**, 1017–1029.