Illuminating the druggable genome through patent bioactivity data

Maria P Magariños [1], **Anna Gaulton** [1, 2], **Eloy Félix** [1], **Tevfik Kiziloren** [1], **Ricardo Arcila** [1], **Tudor I Oprea** [3], **Andrew R Leach** [Corresp. 1]

[1] EMBL-EBI, Hinxton, United Kingdom

[2] Exscientia, Oxford, United Kingdom

[3] Translational informatics Division, Department of Internal Medicine, School of Medicine, University of New Mexico, Albuquerque, United States

Corresponding Author: Andrew R Leach
Email address: arl@ebi.ac.uk

The patent literature is a potentially valuable source of bioactivity data. In this paper we describe a process to prioritise 3.7 million life science relevant patents obtained from the SureChEMBL database (https://www.surechembl.org/), according to how likely they were to contain bioactivity data for potent small molecules on less-studied targets, based on the classification developed by the Illuminating the Druggable Genome (IDG) project. The overall goal was to select a smaller number of patents that could be manually curated and incorporated into the ChEMBL database. Using relatively simple annotation and filtering pipelines, we have been able to identify a substantial number of patents containing quantitative bioactivity data for understudied targets that had not previously been reported in the peer-reviewed medicinal chemistry literature. We quantify the added value of such methods in terms of the numbers of targets that are so identified, and provide some specific illustrative examples. Our work underlines the potential value in searching the patent corpus in addition to the more traditional peer-reviewed literature. The small molecules found in these patents, together with their measured activity against the targets, are now accessible via the ChEMBL database.

Illuminating the Druggable Genome through Patent Bioactivity Data

María Paula Magariños [1], Anna Gaulton [1,2], Eloy Félix [1], Tevfik Kiziloren [1], Ricardo Arcila [1], Tudor I. Oprea [3], Andrew R. Leach [1]

[1] European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK
[2] Current address: Exscientia, The Schrödinger Building, Heatley Road, Oxford Science Park, Oxford, OX4 4GE, UK
[3] Department of Internal Medicine, University of New Mexico School of Medicine, Albuquerque, NM, USA

Corresponding Author:
Andrew R. Leach [1]
European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire CB10 1SD, UK
Email address: arl@ebi.ac.uk

Abstract

The patent literature is a potentially valuable source of bioactivity data. In this paper we describe a process to prioritise 3.7 million life science relevant patents obtained from the SureChEMBL database (https://www.surechembl.org/), according to how likely they were to contain bioactivity data for potent small molecules on less-studied targets, based on the classification developed by the Illuminating the Druggable Genome (IDG) project. The overall goal was to select a smaller number of patents that could be manually curated and incorporated into the ChEMBL database. Using relatively simple annotation and filtering pipelines, we have been able to identify a substantial number of patents containing quantitative bioactivity data for understudied targets that had not previously been reported in the peer-reviewed medicinal chemistry literature. We quantify the added value of such methods in terms of the numbers of targets that are so identified, and provide some specific illustrative examples. Our work underlines the potential value in searching the patent corpus in addition to the more traditional peer-reviewed literature. The small molecules found in these patents, together with their measured activity against the targets, are now accessible via the ChEMBL database.

Introduction

One of the most useful and compelling pieces of evidence for the druggability of a new biological target is the existence of molecules that bind with sufficient affinity to modulate the

biological activity of the target. However, only about 11 % of the proteome has either an approved drug or a compound known to modulate it (Sheils et al., 2021). Chemical probes represent a special type of small molecule for use in target validation studies, not only having good activity against the target but also selectivity, cellular activity, and potentially other relevant criteria (Workman & Collins, 2010; Garbaccio & Parmee, 2016) and are subjected to a peer review process to ensure the quality of any conclusions when used by the wider community. The availability of open-access, public databases such as ChEMBL (Mendez et al., 2019) has greatly simplified the task of identifying potential molecules by providing easy access to more than 19 million bioactivity data points on almost 2 million compounds. The source of these data is primarily the peer-reviewed scientific literature manually extracted by curators; but some of the data has been integrated from other databases including PubChem BioAssay (Kim et al., 2021) and BindingDB (Gilson et al., 2016), and data is also deposited directly from experimental groups.

An additional and potentially valuable source of information and data on bioactive molecules is the patent literature. In drug discovery, patents are routinely filed to protect novel inventions, by both industrial organisations and academic institutions. The relationship between the patent and the "traditional" (academic journal, peer-reviewed) literature has been examined in various published studies, mostly focussed on key questions relating to overlap/duplication and publication date. For example, in a 2009 paper the authors found that just 6% of compounds in patents also appeared in the scientific literature in one of the commercial sources included in their study (GVKBIO) (Southan, Varkonyi & Muresan, 2009). A later study examined 130 drug-target pairs and on average found them published in patents 3.7 years earlier than in scientific papers (Senger, 2017). A 2017 study concluded that the first molecules for a novel target are more likely to be published first in the literature, whereas novel small molecules more frequently appear first in patents than in literature, regardless of which targets they modulate (Ashenden et al., 2017). Finally, a more recent study selected medicinal patents published between 2014 and 2019 and identified patents with information on small molecules, antibodies and vaccines that could potentially be repurposed for cancer related therapies. Some of the drug-disease links found were not present in scientific literature, while others were found in the papers; in some cases these were published before the patents and in others afterwards (Mucke, 2021).

These and similar studies suggest that the patent corpus potentially represents a wealth of information that is not available elsewhere and/or may appear in the scientific literature only after a significant time delay.

Previous studies that have attempted to search or annotate pharmaceutical patents with target-compound information include Akhondi et al. who produced a set of 198 patents manually annotated with chemical compounds, diseases, targets and modes of action by four different groups of curators (Akhondi et al., 2014); Suriyawongkul et al. who tried to identify targets in titles, abstracts and claims of patents that contained bioactive compound information, combining the search for target names with the search for some keywords that were related to bioactivity data (Suriyawongkul, Southan & Muresan, 2010); Tyrchan et al. who compared different methods to extract the key compound from a given patent, and then applied one of the methods to inform the design of AXL kinase inhibitors (Tyrchan et al., 2012); Fechete et al. who searched full-text patents using keywords related to diabetic nephropathy and further narrowed the search

by rules related to frequency and/or patent section of the keywords found, and subsequently extracted the genes mentioned in the claims section of these patents (Fechete et al., 2011); Gigani et al. who performed a search in the SureChEMBL database (Papadatos et al., 2016) using keyword and/or chemical structure searches, with the goal of identifying patents with compounds that could activate the BK_{Ca} channel (Gigani et al., 2016); and Gadiya et al. who developed a tool (PEMT) to identify patents using genes as a starting point, searching for compounds in ChEMBL with activity data for each gene and then searching SureChEMBL using the compounds found in ChEMBL (Gadiya et al., 2022). These, and similar studies, also confirm that extracting information from patents poses many challenges, given the length and complexity of these documents.

Patent data are currently freely available from a number of resources, including Google Patents (https://patents.google.com/), The Lens (https://www.lens.org/), Espacenet (https://www.epo.org/searching-for-patents/technical/espacenet.html), Patentscope (https://patentscope.wipo.int/search/en/search.jsf) and Free Patents Online (https://www.freepatentsonline.com/). All of these systems allow searching for patents using various criteria. Pubchem provides links to the Patentscope database from the World Intellectual Property Organization (WIPO) for more than 16 million compounds, which allows users to find the patents associated with each of these chemical structures (Kim et al., 2021). BindingDB (Gilson et al., 2016) includes a curated set of US granted patents, from which protein-compound activity data is extracted.

In the work reported here, we use SureChEMBL (Papadatos et al., 2016) (https://www.surechembl.org/), which is a fully automated, chemical-structure-enabled database providing the research community with open and free access to the patent literature. Currently, SureChEMBL sources data from both patent applications and granted patents via full text patents from the United States Patent and Trademark Office (USPTO), the European Patent Office (EPO) and the World Intellectual Property Organization (WIPO), and titles and abstracts from the Japanese Patent Office (JPO).

SureChEMBL currently contains ~140 million patents with ~50,000 added monthly. Of these, ~25 million patents are chemically annotated. Approximately 80,000 new compounds are extracted and added each month to the SureChEMBL chemistry database which currently contains more than 25 million unique structures. The pipeline for the extraction of chemical compounds from patents has been described in detail (Papadatos et al., 2016). In summary, chemical entity names, images and molfiles associated with each patent are converted into chemical structures and then registered into a structure-searchable database. This process is fully automated, without requiring any manual step or curation. The data in SureChEMBL can be accessed via a web interface that enables users to perform text and chemical structure queries, filter the output and then display the results. The complete set of chemical structures and patent associations is also available for download.

The US National Institutes of Health established the Illuminating the Druggable Genome (henceforth IDG) project in 2014 (https://commonfund.nih.gov/idg), with the goal of increasing the knowledge about understudied proteins that belong to well-studied protein families, such as ion channels, G-protein coupled receptors (GPCR) and protein kinases. One of the key

deliverables of the IDG project is an informatics platform, Pharos (https://pharos.nih.gov/), that provides researchers with free access to relevant data on targets. An important aspect of the IDG project (and the data in Pharos) is the classification of human proteins into four Target Development Level (TDL) families, based on how well studied these proteins are. In the Tclin category are targets of at least one approved drug; Tchem targets do not have approved drugs but are modulated by at least one small molecule with a potency above the cut-off specified for the target protein family (≤ 30 nM for kinases, ≤ 100 nM for GPCRs and nuclear receptors, ≤ 10 μM for ion channels and ≤ 1 μM for other target families); Tbio targets do not have chemistry qualifying for the Tclin/Tchem categories but satisfy the criteria described at http://juniper.health.unm.edu/tcrd/; while Tdark targets are understudied proteins with little annotation (Oprea et al., 2018). Of particular relevance to the work here is the availability of small molecule modulators for new targets, consistent with other work suggesting that the lack of high-quality chemical probes for understudied targets is an important cause for lack of interest (Edwards et al., 2011; Oprea et al., 2018). The default IDG process uses bioactivity data from ChEMBL (Mendez et al., 2019), as well as from Guide to Pharmacology, which contains manually curated information on ligands and drug targets (Armstrong et al., 2020), and DrugCentral (Avram et al., 2021), which contains bioactivity data annotated from a variety of sources, including the scientific literature. DrugCentral has also information on drugs that have been approved by the United States Food and Drug Administration, the European Medicines Agency, and the Pharmaceuticals and Medical Devices Agency in Japan. The DrugCentral drug information is used in the IDG workflow to identify the targets that belong to the Tclin category (TCRD Home Page, http://juniper.health.unm.edu/tcrd/). At the time of writing, Tclin proteins represent ~3 % of the human proteome; Tchem proteins represent ~8 %; Tbio proteins represent ~58 %; and Tdark proteins represent ~31 % (Sheils et al., 2021).

In this paper, we describe methods to systematically mine the SureChEMBL patent corpus to identify new bioactivity data for Tdark/Tbio targets, with the aim of 1) including the bioactivity data in the ChEMBL database and 2) promoting some of these targets to IDG Tchem status.

Methods

Patents were processed using perl scripts written for this project, accessible via a GitHub repository (https://github.com/chembl/idg_patents_paper).

The starting point for our work was the set of patents extracted from SureChEMBL covering the years 2012 to 2018, flagged as life-science relevant according to the International Patent Classification (IPC) (https://www.wipo.int/classifications/ipc/en/) or the Cooperative Patent Classification (CPC) codes (https://worldwide.espacenet.com/classification) present in the patents. These codes classify the patents into different areas of technology. The specific codes taken into account by the life science flag are: A01, A23, A24, A61, A62B, C05, C06, C07, C08, C09, C10, C11, C12, C13, C14, G01N, which cover a broader set of patents than required but is still useful to filter out many patents that would not be relevant. This resulted in a set of 3.7 million patents.

The goal was to find patents with bioactivity data on small molecules against understudied targets (Tdark or Tbio categories according to the IDG classification). Firstly, in order to determine which patents were likely to have bioactivity data, the files corresponding to the patents were processed to identify tables containing the following keywords: IC50, XC50, EC50, AC50, Ki, Kd, pIC50, pXC50, pEC50, pAC50, -log(IC50), -log(XC50), -log(EC50), -log(AC50), concentration to inhibit, IC-50, XC-50, EC-50, AC-50, IC 50, XC 50, EC 50, AC 50. Out of the 3.7 million patents, 69,289 patents (2 %) were thus flagged as potentially containing bioactivity data in tables (for simplicity called "patents with bioactivity tables").

Separately, we identified patents that might contain information about IDG Tbio and Tdark targets. A list of Tdark and Tbio IDG target names and gene symbols was obtained from the Target Central Resource Database (36,044 target names/symbols) (TCRD Home Page, http://juniper.health.unm.edu/tcrd/). We searched for these target names and their gene symbols in the patent titles, abstracts, descriptions and claims sections, in the context of specific phrases that could indicate bioactivity data of small molecules against them:

- X inhibitors
- Inhibitors of X
- X inhibitor
- Modulators of X
- Modulation of X
- Targeting X
- X modulators
- Binding specifically to X
- X mutants
- Inhibit X
- Antibodies recognis|zing X
- Modulating the X
- Selective X inhibitors
- X antagonists
- X agonist
- X selective binding compounds
- Activity of X
- X antibodies
- X activity
- Inhibitor of X
- X binding
- Antibodies directed against X
- Treatment of X related
- Antibody for X
- Anti-X antibody
- Human anti-X
- Antibodies to X
- High X affinity
- Inhibiting X

- Blocks|block X
- Blocking X
- Ligand|ligands for X
- Compounds that interact with X
- Modulating the function of X
- X ligand|ligands

The combination of these two procedures allowed us to classify the patents into six groups: patents with bioactivity tables, and targets mentioned in titles or abstracts; patents with bioactivity tables, and targets mentioned in descriptions or claims sections; patents without bioactivity tables, and targets mentioned in titles or abstracts; patents without bioactivity tables, and targets mentioned in descriptions and claims; patents with bioactivity tables but no targets; and patents without bioactivity tables and without targets (Fig. 1). This was done with the goal of prioritising the patents, with the expectation that most data would be found in Group 1, followed by Group 2; we expected Groups 3 and 4 to contain fewer patents with bioactivity data (given that they were not flagged as containing bioactivity tables). The patents in Groups 5 and 6 did not have target matches and for this reason were not expected to have bioactivity data against the understudied targets of interest to us.

Following this automated annotation/filtering process, a number of patents from each group were manually examined to confirm the presence of the correct Tbio/Tdark target, the presence of quantitative bioactivity measurements, and that the Tbio/Tdark target was the molecular target to which these bioactivity measurements applied. This final check is required because some of the patents were found to have data only on targets that did not belong to the IDG list of understudied targets; other patents did contain data exclusively on the targets of particular interest to us. Some patents fell into both categories.

For patents with confirmed bioactivity data, details of compounds synthesised, biological assays performed, and bioactivity measurements were manually extracted according to the standard ChEMBL curation procedure described previously (Gaulton et al., 2015) and loaded into the ChEMBL database. Briefly, structures and names of all tested compounds were extracted, together with a description of the assays performed, name of the targets, species, and measurement values and units. Compound structures were standardised and integrated into ChEMBL, mapping them to an existing structure or creating a new entry in the database as appropriate.

In addition to registering the reported measurement values, the bioactivity data obtained was standardised to facilitate comparison of results for common activity types. Bioactivity data was also mapped to existing ChEMBL targets according to species and sequence or accession. When this was not possible a new target was created and then mapped to the corresponding assay.

All bioactivities against all the targets present in these patents (irrespective of their inclusion or not in the IDG Tbio/Tdark categories) were extracted by the curators.

Results

As a result of this work, bioactivity data from 225 patents were loaded into ChEMBL, corresponding to 657 targets (including single proteins, protein families, protein complexes, organisms, cell lines and protein-protein interactions) and 18,319 compounds. For 145 of these targets, this represents the only source of information of bioactivity data in ChEMBL.

A patent family is a set of patents of identical content (European Patent Office, https://www.epo.org/searching-for-patents/helpful-resources/first-time-here/patent-families/docdb.html). The scripts described here were run against every patent in SureChEMBL that belong to the set of 2012-2018 patents flagged as life science related. In order to avoid duplication of effort, patents were grouped by patent family. For this reason, the patent counts in the sections below are given as number of patent families rather than number of patents.

We examined the distribution of positive and negative patent families among the different groups shown in Fig. 1, to identify which group or groups were more or less likely to contain useful information, as this might facilitate the task of identifying the most useful patents for future analyses. The group that had the highest percentage of positive patents was Group 1: 49 positive patent families in the 291 families examined (16.8 %), followed by Group 3: 88 positive families in the 1,912 examined (4.6 %). There was 1 patent family in common between the positive patents of these two groups. Group 2 had 92 positive patent families, but 46 of them were already present in Group 1. Group 4 had 96 positive patent families, but 86 of them were already present in Group 3. There were very few patents with data in Group 5 or Group 6 (0.1 % and 0.4 % patent families of the examined ones, respectively) (Fig. 1).

A full list of patents and the targets they contain can be found in Table S1. Note that one patent is omitted from this list (US-8409550-B2) because it contains data against a target from *Bos taurus*, whereas IDG is focussed solely on human targets. 76 of these targets had at least one compound with bioactivity data values within the cut-off for its target family, as defined by the target class-specific IDG criteria outlined earlier. Table S2 shows which targets had bioactivity data within the cut-offs for its target family, and Table S3 shows how many patents, total compounds and compounds within the cut-off were found for each IDG target class.

As BindingDB also extracts data from patents, we were interested in examining the overlap between the two data sets. For all the targets found, we performed a search by target name in BindingDB with the goal of comparing the results from the two different databases. Because BindingDB extracts only US granted patents, we used the patent family to do the comparison. We found 33 targets in both BindingDB and the dataset from our method. Of the 70 patent families found by our method for these targets, 20 were also found in BindingDB. 50 families were found exclusively with our method, and 34 families were found exclusively by BindingDB. In most cases, the patents that were missed had targets mentioned using a name that was not on our list of targets to find (for example, “CH24H” instead of “Cholesterol 24-hydroxylase”). In other cases the patents belonged to Groups 3 or 4 and were not part of the set of patents that we selected to read.

Examples of understudied targets with bioactivities found in patents

In this section we briefly describe three specific examples of targets for which we were able to identify and curate bioactivity data from the patent workflow described above. Some of the compounds found for each target are shown in Fig. 2.

LATS1

LATS1 is a Ser/Thr kinase that belongs to the LATS (large tumor suppressor) family (Xu et al., 2015). It is a component of the Hipo pathway which is involved in cancer, organ development, growth and regeneration (Fu, Plouffe & Guan, 2017), and cell contact inhibition (Zheng & Pan, 2019).
This kinase is conserved among several organisms, such as yeast, nematodes, flies, and mammals. In humans, LATS1 can be found in high levels in most tissues, and it has a role in regulation of mitosis and apoptosis. In some types of cancer there is evidence of mutations in LATS1, and of LATS1 inactivation through promoter hypermethylation in others (Furth & Aylon, 2017).
It was found that the expression of this protein is elevated in some cancers, but decreased in others (Xu et al., 2015; Furth & Aylon, 2017).
Without considering data from patents, there are currently 430 molecules in ChEMBL associated with this target, extracted from 44 different papers, but only four molecules satisfy the IDG cut-off criteria for kinases. These molecules were extracted from four different papers: PMID 19654408 (Zarrinkar et al., 2009), PMID 22037378 (Davis et al., 2011), PMID 29191878 (Klaeger et al., 2017) and PMID 30384048 (Narayan et al., 2019). As a result of the current work, there are now 289 additional molecules associated with this target, with 184 molecules satisfying the IDG cut-off criteria to promote the target to the Tchem category. The source of these compounds is the patent US-20180344738-A1 (Behnke et al., 2018), which describes molecules designed to promote cell proliferation, with applications such as chronic wound healing, promoting liver regrowth, or treating limbal stem cell deficiency.

Histone-lysine N-methyltransferase SUV39H2

SUV39H2 is a lysine methyltransferase, first identified in *Drosophila*, which methylates histone 3 on lysine 9 (H3K9). Di- and trimethylation of H3K9 results in gene expression repression (Kaniskan, Martini & Jin, 2018).
This protein is present only in embryogenesis and adult testis of healthy individuals (Li, Zheng & Yang, 2019), but overexpressed in several cancers, for example lung adenocarcinoma, colorectal carcinoma and gastric carcinoma (Saha & Muntean, 2021).
There are no selective inhibitors for this target (Kaniskan, Martini & Jin, 2018).
At the start of this work there were 19 molecules in ChEMBL with bioactivity data against SUV39H2, all of them from scientific literature, but none of these molecules were within the IDG cut-off.

Our patent workflow identified 460 molecules from just a single patent (US-20180273529-A1) (Matsuo et al., 2018), all within the corresponding IDG cut-off.

G protein-coupled receptor 6

GPR6 is a G-protein coupled receptor, still classified as orphan by the International Union of Basic and Clinical Pharmacology (IUPHAR) (Alexander et al., 2019) due to lack of consistency among reports related to endogenous ligands (Morales, Isawi & Reggio, 2018; Laun et al., 2019). It is expressed mainly in neurons in mammalian striatum and hypothalamus. There is evidence that it could have a role in several processes and diseases such as neurite outgrowth, instrumental learning, Alzheimer's disease, Parkinson's disease, Huntington's disease (Morales, Isawi & Reggio, 2018; Laun et al., 2019), and schizophrenia (Morales, Isawi & Reggio, 2018).
At the start of this work, there were 227 molecules in ChEMBL with bioactivity data values within the IDG cut-off. These molecules were obtained from patents, either from BindingDB, or our own curation efforts previous to the work described here.
As a result of this search, 100 additional molecules with bioactivity against GPR6 with values within the cut-off of ≤ 100 nM were identified, from patents WO-2018183145-A1 (Green et al., 2018) and EP-2882722-A1 (Hitchcock et al., 2015). Figure 3 shows a timeline with patent and scientific literature numbers by year for GPR6, showing that in this particular case, significantly more data were reported via patent disclosures than in the scientific literature. Interestingly, a new clinical candidate (currently in phase 2 clinical trials) for Parkinson's disease, CVN424 (a GPR6 inverse agonist), has been disclosed (Sun et al., 2021; Brice et al., 2021). This molecule can be found in patents as early as 2015 in patent US-9181249-B2 (Brown et al., 2015).

Discussion

The overall goal of this work was to identify bioactivity data on understudied targets from the patent literature, which could allow us to promote targets to the IDG Tchem category. We focused on small molecules only, but our workflow also identified several patents concerning antibodies or RNA as therapeutic agents. For the purposes of this work, we did not progress these patents further, but clearly they could also be useful in the context of "illuminating" new targets.
It should be noted that the work described here was conducted over a period of time, during which complementary data from other sources was added to the various resources concerned. This reflects the natural evolution of the underlying databases, each with their own update mechanism and release schedule procedures. Thus, for example, targets designated as Tdark or Tbio at the time when the research was initiated may have been promoted to a higher category based on separate evidence whilst the patent bioactivity work described here was underway. In the narrative below, the data reflects the particular snapshot corresponding to the time on which the work was initiated or completed, as appropriate.
Two of the proteins included in the list of targets that we searched for (sclerostin and exportin-1) were originally classified as Tbio at the start of this work, but later promoted to Tclin after the approval of the first-in-class drugs romosozumab and selinexor respectively.

Coincidentally, over the same period of time, bioactivity data for 30 Tbio/Tdark targets were added into ChEMBL from the scientific literature. There was an overlap of 21 targets between the two sets. This shows the value of using patents as an additional source of bioactivity data. The work described here involved manually reviewing many patents from the six groups described in Fig. 1. As expected, the group with higher percentage of positive patents was Group 1, unexpectedly followed by the patents in Group 3, which even though they could not be flagged as containing bioactivity tables, still contained bioactivity data was not detected automatically with the method used here, and were only found when reviewing the patent manually. Groups 2 and 4 had lower percentages but still delivered useful and relevant patents. Even though classifying the patents in this way provided a starting point for prioritisation, clearly there some limitations to this approach as shown in the comparison with BindingDB and the reasons for missing some patents, and for future work it would be advantageous to develop methods that can reduce the level of manual review that is required. This is the focus of currently ongoing work to develop machine-learning models able to predict which patents should be prioritised for human examination and potential curation.
In addition to the two targets that are now Tclin, 74 Tdark/Tbio targets were promoted to the Tchem category on the basis of the bioactivity data identified from our patent analysis.

Conclusions

Using relatively simple annotation and filtering pipelines, we have been able to identify a substantial number of patents containing quantitative bioactivity data for understudied targets that had not previously been reported in the peer-reviewed medicinal chemistry literature. This underlines the potential value in searching the patent corpus in addition to the more traditional peer-reviewed literature. The small molecules found in these patents, together with their measured activity against the targets, are now accessible via the ChEMBL database and Pharos, and have contributed to the “illumination” of previously dark targets.

Acknowledgements

We would like to thank Dr Barbara Zdrazil for helpful comments on this manuscript.

References

Akhondi SA, Klenner AG, Tyrchan C, Manchala AK, Boppana K, Lowe D, Zimmermann M, Jagarlapudi SA, Sayle R, Kors JA, and Muresan S. 2014. Annotated chemical patent corpus: a gold standard for text mining. PLoS One 9:e107477. 10.1371/journal.pone.0107477

Alexander SP, Battey J, Benson HE, Benya RV, Bonner TI, Davenport AP, Eguchi S, Harmar A, Holliday N, Jensen RT, Karnik S, Kostenis E, Liew WC, Monaghan AE, Mpamhanga C, Neubig R, Pawson AJ, Pin J-P, Sharman JL, Spedding M, Spindel E, Stoddart L, Storjohann L, Thomas WG, Tirupula K, and Vanderheyden P. 2019. Class A Orphans (version 2019.5) in the IUPHAR/BPS Guide to Pharmacology Database. IUPHAR/BPS Guide to Pharmacology CITE, 2019(5). doi: 10.2218/gtopdb/F16/2019.5

Armstrong JF, Faccenda E, Harding SD, Pawson AJ, Southan C, Sharman JL, Campo B, Cavanagh DR, Alexander SPH, Davenport AP, Spedding M, Davies JA, and Nc I. 2020. The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY. Nucleic Acids Res 48:D1006-D1021. 10.1093/nar/gkz951

Ashenden SK, Kogej T, Engkvist O, and Bender A. 2017. Innovation in Small-Molecule-Druggable Chemical Space: Where are the Initial Modulators of New Targets Published? J Chem Inf Model 57:2741-2753. 10.1021/acs.jcim.7b00295

Avram S, Bologa CG, Holmes J, Bocci G, Wilson TB, Nguyen DT, Curpan R, Halip L, Bora A, Yang JJ, Knockel J, Sirimulla S, Ursu O, and Oprea TI. 2021. DrugCentral 2021 supports drug discovery and repositioning. Nucleic Acids Res 49:D1160-D1169. 10.1093/nar/gkaa997

Behnke D, Berenshteyn F, Hao X, Hoffman T, Jin Q, Lacoste A, Lee C, Liu J, Liu Y, Maibaum JK, Mo T, Pan J, Qu X, Tchorz J, Xie YF, Yan S, Zou Y. 2018. 6-6 Fused Bicyclic Heteroaryl Compounds and their Use as LATS Inhibitors. Patent number US-20180344738-A1.

Brice NL, Schiffer HH, Monenschein H, Mulligan VJ, Page K, Powell J, Xu X, Cheung T, Burley JR, Sun H, Dickson L, Murphy ST, Kaushal N, Sheardown S, Lawrence J, Chen Y, Bartkowski D, Kanta A, Russo J, Hosea N, Dawson LA, Hitchcock SH, and Carlton MB. 2021. Development of CVN424: A Selective and Novel GPR6 Inverse Agonist Effective in Models of Parkinson Disease. J Pharmacol Exp Ther 377:407-416. 10.1124/jpet.120.000438

Davis MI, Hunt JP, Herrgard S, Ciceri P, Wodicka LM, Pallares G, Hocker M, Treiber DK, and Zarrinkar PP. 2011. Comprehensive analysis of kinase inhibitor selectivity. Nat Biotechnol 29:1046-1051. 10.1038/nbt.1990

Brown J, Hitchcock S, Hopkins M, Kikuchi S, Monenschein H, Reichard H, Schleicher K, Sun H, Macklin T. 2015. Substituted pyrido[3,4-b]pyrazines as GPR6 modulators. Patent number US-9181249-B2.

Edwards AM, Isserlin R, Bader GD, Frye SV, Willson TM, and Yu FH. 2011. Too many roads not taken. Nature 470:163-165. 10.1038/470163a

European Patent Office. https://www.epo.org/searching-for-patents/helpful-resources/first-time-here/patent-families/docdb.html. Accessed 12 of January 2023.

Fechete R, Heinzel A, Perco P, Monks K, Sollner J, Stelzer G, Eder S, Lancet D, Oberbauer R, Mayer G, and Mayer B. 2011. Mapping of molecular pathways, biomarkers and drug targets for diabetic nephropathy. Proteomics Clin Appl 5:354-366. 10.1002/prca.201000136

Fu V, Plouffe SW, and Guan KL. 2017. The Hippo pathway in organ development, homeostasis, and regeneration. Curr Opin Cell Biol 49:99-107. 10.1016/j.ceb.2017.12.012

Furth N, and Aylon Y. 2017. The LATS1 and LATS2 tumor suppressors: beyond the Hippo pathway. Cell Death Differ 24:1488-1501. 10.1038/cdd.2017.99

Gadiya Y, Zaliani A, Gribbon P, and Hofmann-Apitius M. 2022. PEMT: a patent enrichment tool for drug discovery. Bioinformatics. 10.1093/bioinformatics/btac716

Garbaccio RM, and Parmee ER. 2016. The Impact of Chemical Probes in Drug Discovery: A Pharmaceutical Industry Perspective. Cell Chem Biol 23:10-17. 10.1016/j.chembiol.2015.11.011

Gaulton A, Kale N, van Westen GJ, Bellis LJ, Bento AP, Davies M, Hersey A, Papadatos G, Forster M, Wege P, and Overington JP. 2015. A large-scale crop protection bioassay data set. Sci Data 2:150032. 10.1038/sdata.2015.32

Gigani Y, Gupta S, Lynn A, and Asotra K. 2016. Identification of BKCa channel openers by molecular field alignment and patent data-driven analysis. Pharmaceutical and Biomedical Research 2:22-29. 10.18869/acadpub.pbr.2.4.22

Gilson MK, Liu T, Baitaluk M, Nicola G, Hwang L, and Chong J. 2016. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res 44:D1045-1053. 10.1093/nar/gkv1072

Green J, Hopkins M, Jones B, Kiryanov AA, Kuehler J, Monenschein H, Murphy S, Nixey T, Sun H. 2018. Piperidinyl- and piperazinyl-substituted heteroaromatic carboxamides as modulators of GPR6. Patent number WO-2018183145-A1.

Hitchcock S, Monenschein H, Reichard H, Sun H, Kikuchi S, Macklin T, Hopkins M. 2015. Quinoxaline derivatives as GPR6 modulators. Patent number EP-2882722-A1.

Kaniskan HU, Martini ML, and Jin J. 2018. Inhibitors of Protein Methyltransferases and Demethylases. Chem Rev 118:989-1068. 10.1021/acs.chemrev.6b00801

Kim S, Chen J, Cheng T, Gindulyte A, He J, He S, Li Q, Shoemaker BA, Thiessen PA, Yu B, Zaslavsky L, Zhang J, and Bolton EE. 2021. PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 49:D1388-D1395. 10.1093/nar/gkaa971

Klaeger S, Heinzlmeir S, Wilhelm M, Polzer H, Vick B, Koenig PA, Reinecke M, Ruprecht B, Petzoldt S, Meng C, Zecha J, Reiter K, Qiao H, Helm D, Koch H, Schoof M, Canevari G, Casale E, Depaolini SR, Feuchtinger A, Wu Z, Schmidt T, Rueckert L, Becker W, Huenges J, Garz AK, Gohlke BO, Zolg DP, Kayser G, Vooder T, Preissner R, Hahne H, Tonisson N, Kramer K, Gotze K, Bassermann F, Schlegl J, Ehrlich HC, Aiche S, Walch A, Greif PA, Schneider S, Felder ER, Ruland J, Medard G, Jeremias I, Spiekermann K, and Kuster B. 2017. The target landscape of clinical kinase drugs. Science 358. 10.1126/science.aan4368

Laun AS, Shrader SH, Brown KJ, and Song ZH. 2019. GPR3, GPR6, and GPR12 as novel molecular targets: their biological functions and interaction with cannabidiol. Acta Pharmacol Sin 40:300-308. 10.1038/s41401-018-0031-9

Li B, Zheng Y, and Yang L. 2019. The Oncogenic Potential of SUV39H2: A Comprehensive and Perspective View. J Cancer 10:721-729. 10.7150/jca.28254

Matsuo Y, Hisada S, Nakamura Y, Chakrabarti A, Rawat M, Rai S, Satyanarayana AV, Duan Z, Talukdar A, Ravula S, Decornez H. 2018. Bicyclic compound and use thereof for inhibiting SUV39H2. Patent number US-20180273529-A1.

Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Felix E, Magarinos MP, Mosquera JF, Mutowo P, Nowotka M, Gordillo-Maranon M, Hunter F, Junco L, Mugumbate G, Rodriguez-Lopez M, Atkinson F, Bosc N, Radoux CJ, Segura-Cabrera A, Hersey A, and Leach AR. 2019. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47:D930-D940. 10.1093/nar/gky1075

Morales P, Isawi I, and Reggio PH. 2018. Towards a better understanding of the cannabinoid-related orphan receptors GPR3, GPR6, and GPR12. Drug Metab Rev 50:74-93. 10.1080/03602532.2018.1428616

Mucke HAM. 2021. What patents tell us about drug repurposing for cancer: A landscape analysis. Semin Cancer Biol 68:3-7. 10.1016/j.semcancer.2019.09.010

Narayan S, Ramisetti S, Jaiswal AS, Law BK, Singh-Pillay A, Singh P, Amin S, and Sharma AK. 2019. ASR352, A potent anticancer agent: Synthesis, preliminary SAR, and biological activities against colorectal cancer bulk, 5-fluorouracil/oxaliplatin resistant and stem cells. Eur J Med Chem 161:456-467. 10.1016/j.ejmech.2018.10.052

Oprea TI, Bologa CG, Brunak S, Campbell A, Gan GN, Gaulton A, Gomez SM, Guha R, Hersey A, Holmes J, Jadhav A, Jensen LJ, Johnson GL, Karlson A, Leach AR, Ma'ayan A, Malovannaya A, Mani S, Mathias SL, McManus MT, Meehan TF, von Mering C, Muthas D, Nguyen DT, Overington JP, Papadatos G, Qin J, Reich C, Roth BL, Schurer SC, Simeonov A, Sklar LA, Southall N, Tomita S, Tudose I, Ursu O, Vidovic D, Waller A, Westergaard D, Yang JJ, and Zahoranszky-Kohalmi G. 2018. Unexplored therapeutic opportunities in the human genome. Nat Rev Drug Discov 17:317-332. 10.1038/nrd.2018.14

Papadatos G, Davies M, Dedman N, Chambers J, Gaulton A, Siddle J, Koks R, Irvine SA, Pettersson J, Goncharoff N, Hersey A, and Overington JP. 2016. SureChEMBL: a large-scale, chemically annotated patent document database. Nucleic Acids Res 44:D1220-1228. 10.1093/nar/gkv1253

Saha N, and Muntean AG. 2021. Insight into the multi-faceted role of the SUV family of H3K9 methyltransferases in carcinogenesis and cancer progression. Biochim Biophys Acta Rev Cancer 1875:188498. 10.1016/j.bbcan.2020.188498

Senger S. 2017. Assessment of the significance of patent-derived information for the early identification of compound-target interaction hypotheses. J Cheminform 9:26. 10.1186/s13321-017-0214-2

Sheils TK, Mathias SL, Kelleher KJ, Siramshetty VB, Nguyen DT, Bologa CG, Jensen LJ, Vidovic D, Koleti A, Schurer SC, Waller A, Yang JJ, Holmes J, Bocci G, Southall N, Dharkar P, Mathe E, Simeonov A, and Oprea TI. 2021. TCRD and Pharos 2021: mining the human proteome for disease biology. Nucleic Acids Res 49:D1334-D1346. 10.1093/nar/gkaa993

Southan C, Varkonyi P, and Muresan S. 2009. Quantitative assessment of the expanding complementarity between public and commercial databases of bioactive compounds. J Cheminform 1:10. 10.1186/1758-2946-1-10

Sun H, Monenschein H, Schiffer HH, Reichard HA, Kikuchi S, Hopkins M, Macklin TK, Hitchcock S, Adams M, Green J, Brown J, Murphy ST, Kaushal N, Collia DR, Moore S, Ray WJ, English NM, Carlton MBL, and Brice NL. 2021. First-Time Disclosure of CVN424, a Potent and Selective GPR6 Inverse Agonist for the Treatment of Parkinson's Disease: Discovery, Pharmacological Validation, and Identification of a Clinical Candidate. J Med Chem 64:9875-9890. 10.1021/acs.jmedchem.0c02081

Suriyawongkul I, Southan C, and Muresan S. 2010. The Cinderella of Biological Data Integration: Addressing Some of the Challenges of Entity and Relationship Mining from Patent Sources. Berlin, Heidelberg: Springer Berlin Heidelberg. p 106-121.

TCRD Home Page. http://juniper.health.unm.edu/tcrd/. Accessed 20 of January 2023.

Tyrchan C, Bostrom J, Giordanetto F, Winter J, and Muresan S. 2012. Exploiting structural information in patent specifications for key compound prediction. J Chem Inf Model 52:1480-1489. 10.1021/ci3001293

Workman P, and Collins I. 2010. Probing the probes: fitness factors for small molecule tools. Chem Biol 17:561-577. 10.1016/j.chembiol.2010.05.013

Xu B, Sun D, Wang Z, Weng H, Wu D, Zhang X, Zhou Y, and Hu W. 2015. Expression of LATS family proteins in ovarian tumors and its significance. Hum Pathol 46:858-867. 10.1016/j.humpath.2015.02.012

Zarrinkar PP, Gunawardane RN, Cramer MD, Gardner MF, Brigham D, Belli B, Karaman MW, Pratz KW, Pallares G, Chao Q, Sprankle KG, Patel HK, Levis M, Armstrong RC, James J, and Bhagwat SS. 2009. AC220 is a uniquely potent and selective inhibitor of FLT3 for the treatment of acute myeloid leukemia (AML). Blood 114:2984-2992. 10.1182/blood-2009-05-222034

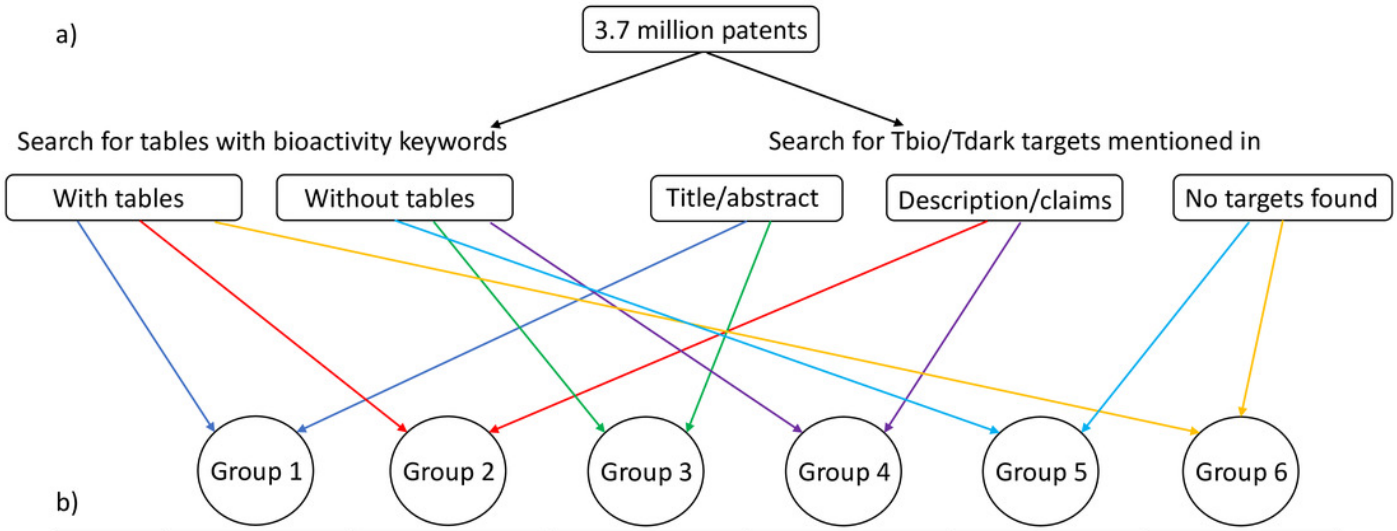
Zheng Y, and Pan D. 2019. The Hippo Signaling Pathway in Development and Disease. Dev Cell 50:264-282. 10.1016/j.devcel.2019.06.003

Figure 1

Patent group classification.

a) 3.7 million patents were scanned for the presence of 1) tables with bioactivity keywords and 2) understudied target names mentioned in the context of specific phrases in the titles, abstracts, description and claims sections of the patents. According to the result of each of these two independent processes, the patents could be classified in the following groups: Group 1: patents with bioactivity tables, and targets mentioned in titles or abstracts; Group 2: patents with bioactivity tables, and targets mentioned in descriptions or claims sections; Group 3: patents without bioactivity tables, and targets mentioned in titles or abstracts; Group 4: patents without bioactivity tables, and targets mentioned in descriptions and claims; Group 5: patents with bioactivity tables but no targets; Group 6: patents without bioactivity tables and without targets. b) A subset of patent families in each group was manually examined. Total: total number of patent families that belonged to each group; Read: number of patent families of each group that were read to determine the presence of relevant data; Positive: number of patent families read that had bioactivity data of small molecules against at least one understudied target; Negative: number of patent families read that did not have bioactivity data of small molecules on understudied targets.

a)



b)

Total	304	5,057	2,128	31,452	1,839,336	23,046
Read	291	3,654	1,912	2,736	989	935
Positive	49	92	88	96	1	4
Negative	242	3,562	1,824	2,640	988	931

Figure 2

Illustrative examples of bioactive compounds identified from SureChEMBL patent workflow against three targets.

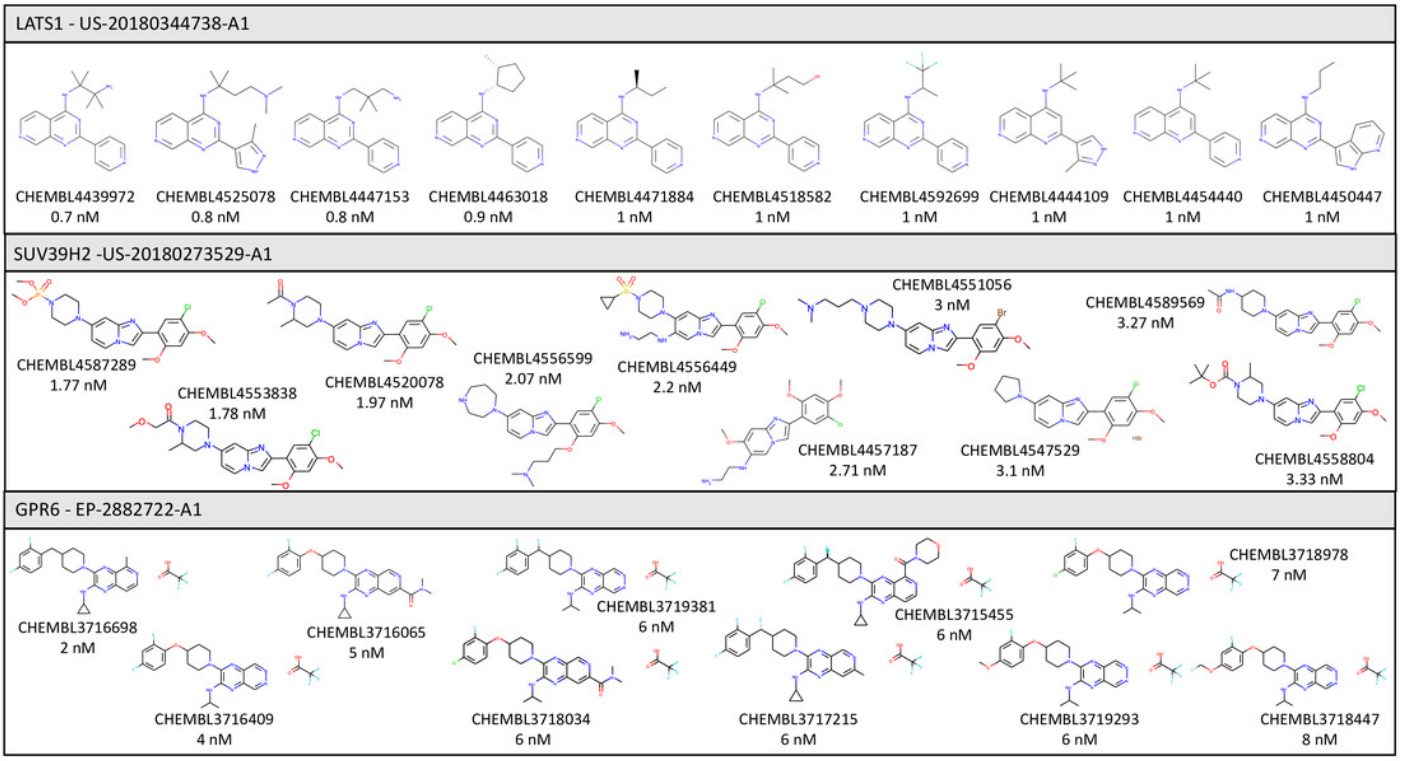


Figure 3

Numbers of relevant patents and scientific literature publications for GPR6 per year.

Key compound disclosures in patents and scientific literature indicated by dashed lines: a) Example of a compound reported in one of the earliest patents with data against GPR6. b) Example of a compound in one of the earliest patents identified by our method. c) First small molecule modulator reported in scientific literature.