Bioinformatics, YYYY, 0–0
doi: xxx/xxx/xxxxx
Advance Access Publication Date: DD Month 2022
Applications Note

Data and Text Mining

PEMT: A patent enrichment tool for drug discovery

Yojana Gadiya[1,2*], Andrea Zaliani[1,2], Philip Gribbon[1,2], Martin Hofmann-Apitius[3,4]

(1) Fraunhofer Institute for Translational Medicine and Pharmacology (ITMP), Schnackenburgallee 114, 22525 Hamburg (2) Fraunhofer Cluster of Excellence for Immune-Mediated Diseases (CIMD), Theodor Stern Kai 7, 60590 Frankfurt, Germany; (3) Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), 53754 Sankt Augustin, Germany; (4) Bonn-Aachen International Center for Information Technology (B-IT), University of Bonn, 53113 Bonn, Germany

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Drug discovery practitioners in industry and academia use semantic tools to extract information from online scientific literature to generate new insights into targets, therapeutics and diseases. However, due to complexities in access and analysis, patent-based literature is often overlooked as a source of information. As drug discovery is a highly competitive field, it is natural that tools tapping into patent literature can provide any actor in the field an advantage in terms of better-informed decision-making. Hence, we aim to facilitate access to patent literature by creating an automatic tool for extracting information from patents described in existing public resources.

Results: Here, we present PEMT, a novel patent enrichment tool, that takes advantage of public databases like ChEMBL and SureChEMBL to extract relevant patent information linked to chemical structures and/or gene names described through FAIR principles and metadata annotations. PEMT is aimed at supporting drug discovery and research in establishing a patent landscape around genes of interest. The pharmaceutical focus of the tool is mainly due to the subselection of IPC codes, but in principle, it can be used for other patent fields, provided that a link between a concept and chemical structure is investigated. Finally, we have demonstrated a use-case in rare diseases by generating a gene-patent list based on the epidemiological prevalence of these diseases and exploring their underlying patent landscape.

Availability and implementation: PEMT is an open-source Python tool and its source code and PyPi package are available at https://github.com/Fraunhofer-ITMP/PEMT and https://pypi.org/project/PEMT/, respectively.

Contact: yojana.gadiya@itmp.fraunhofer.de

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1. Introduction

Patents are a relatively untapped source of scientific information which nevertheless play a vital role in reflecting the progress of organisations involved in drug discovery. The scientific contents of patents can be overlooked as the information is often shifted in time with respect to any corresponding scientific publications due to the lenghty process before public access is finally possible. Moreover, the legal-based language used to describe the claims of granted applications also introduce additional complexities in the data analysis compared to classical scientific publications. However, analysis of patents can be decisive in assessing and reviewing, for instance, a companies' disease strategies (Roskams-Edris *et al.*, 2019) or making well-informed target selection and prioritisation decisions (Jin and Wong, 2014).

Patent literature provides a more realistic view of drug discovery by focusing on industry-specific genes or targets than scientific literature (Mucke, 2021). A patent application is a structured document and patent databases (public and private) usually take advantage of this feature (Donald *et al.*, 2018). Hence, patent resources such as PATENTSCOPE and Espacenet provide patent searching and retrieving capabilities based on drugs or disease names. Despite the intrinsic value in analysing all relevant patent data in a project, there is often a limit on the number of patents individual users can retrieve within any period of time. In addition, many existing resources have associated access fees, which may not be affordable for academia and smaller-scale companies. This situation gives rise to a need for additional open-source resources for analysing patent contents.

Here, we present PEMT, an automated patent extractor tool that takes advantage of information from patent databases and connects genes or chemicals to this information retrospectively. While patent officers are expert professionals, specifically trained in legal aspects of patent analysis, drug researchers have few open-source tools enabling them to perform a qualitative landscape evaluation on genes linked to certain diseases. PEMT is designed to provide them with such a tool. In addition, we confirm the applicability of the tool in the rare disease domain and show how the tool can assist the scientific community in exploring this untapped resource.

2. Methods and Material

2.1. Implementation details

PEMT is written in Python 3.8 and available at PyPi (https://pypi.org/project/PEMT/) and GitHub (https://github.com/Fraunhofer-ITMP/PEMT).

2.2. Data harmonization

PEMT deals with three types of entities: genes, modulators such as chemicals or biological agents, and patents. For each of these entities, a harmonization step occurs where they are mapped into one or multiple well-known identifiers (i.e. genes are normalized to UniProt and ChEMBL IDs, modulators normalized to ChEMBL and SureChEMBL IDs, and patents normalized to the unique

registered global document IDs), thereby increasing alignment with FAIR principles. The reasoning behind the selection of these resources is discussed in **Supplementary Text 1.**

2.3. Design Architecture

As shown in **Figure 1**, the tool takes a two-step approach to collect relevant patent documents. In the first step, extraction of modifiers that directly regulate (i.e. activation or inhibition) specified genes of interest takes place. This is done based on experimental data (bioassay data) that are deposited into public repositories, including ChEMBL. In the second step, interlinking of identified modulators to patent documents is done by systematically querying SureChEMBL, a patent database. Together, these steps aid in connecting patent documents to genes of interest via chemicals or biological modulators.

The tool requires an input of genes in the form of a comma-separated (CSV) or tab-separated (TSV) file with each row representing an individual gene using well-known nomenclature (e.g. HGNC and UniProt). A gene harmonizer is then run on the genes to map them to their corresponding ChEMBL identifiers, which serve as the basis for querying ChEMBL for experimental data with some predefined conditions (**Supplementary Text 2**). With the help of these modulators, PEMT then explores SureChEMBL for patent literature. A chemical harmonization step (where mapping of ChEMBL identifiers to corresponding SureChEMBL identifiers is done) and a patent filtering step (**Supplementary Text 3**) are performed sequentially. With this approach, PEMT can connect genes to the associated patent documents via chemical structure or biological identifiers.

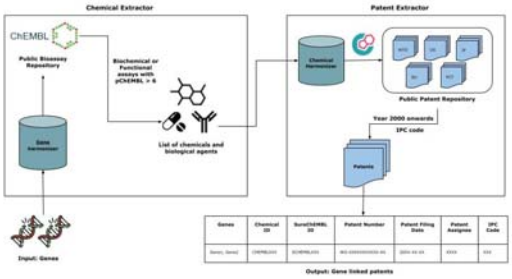


Figure 1. Overview of the framework of the Patent EnrichMent Tool (PEMT).

3. Case scenario

3.1. Data generation

Orphanet (www.orpha.net) is a collection of rare disease data including the Orphanet Ontology of Rare Diseases (ORDO), information on the epidemiological occurrence of rare diseases, as well as the biological mechanisms (disease-gene interactions) involved in rare diseases (Weinreich *et al.*, 2008). We make use of two data categories from Orphanet, i.e. epidemiological data and biological mechanisms, to demonstrate the applicability of PEMT.

As a starting point, we selected a subset of genes based on their corresponding disease prevalence from the current list of 3,886 diseases and 2,637 genes available in Orphanet. The prevalence figure was available from the epidemiological data and diseases that had a prevalence of less than 10 were filtered out. This selection criterion resulted in 59 diseases and 56 genes, representing the most common of the rare diseases, where significant patenting activity would be reasonably anticipated to have occurred in the past. These 56 genes were then provided as inputs for the PEMT tool.

3.2. Data analysis

PEMT generated connections between genes to patents via modulators. Based on the publicly available bioassays, only 12 genes were associated with modulators (**Supplementary Figure 1**). The number of agents ranges between 115 (for FGFR3) and 1 (for MECP2) and a total of 469 were extracted from this step. A further reduction in the chemical space (of the 469 chemicals) was performed, based on their patentability, resulting in 143 modulators. From these 143 moduators, only 97 were identified to be linked to pharmaceutically relevant patents (**Supplementary Table 2**). Ultimately, 135 unique patents belonging to 10 genes (2 did not have any patent information connected to them), linked via 97 modulators, were extracted using the PEMT tool (**Supplementary Figure 2**).

Patents covered multiple geographical jurisdictions: 83 from the United States of America, 19 from the European Patent Office (EPO), and 33 patents from the World Intellectual Property Organisation (WIPO). Furthermore, 106 patents are not yet granted while 29 patents have already been granted (**Supplementary Table 3**). We also found that these patents cover a range of different cooperative patent classification (CPC) codes: C07D (82 patents), C07K (18 patents), C12Q (14 patents), C07H (4 patents), A61P (9 patents), C12P (3 patents), C07F (5 patents), and C07C (1 patent) and are distributed over the year 2000-2021 (**Supplementary Figure 3**). Moreover, upon assessing the patent assignees, we found 85 patents belonged to pharmaceutical industries, 23 patents to academic institutes, and 25 patents to individuals.

Lastly, we looked at the historical timeline for each patented gene to understand the genes' importance over the patent landscape (**Supplementary Figure 4**). It can be observed that patents targeting FGFR3 received high attention in the early 2000s. This gene is associated, among others, with diseases such as Muenke syndrome which has no interventional clinical studies and only a single completed clinical trial in 2005 (NCT00106977) aimed to better understand the disease aetiology. The real attention for FGFR3 in those years was about multiple myeloma and bladder carcinoma (Grand *et al.*, 2004), but in the following years, it demonstrated marginal efficacy in cancer (Chae *et al.*, 2017). On the other hand, targets like SLC2A3 started surfacing in patents in 2015 and contributed to the highest number of patents primarily due to their association with prominent neurodegenerative disease, Huntington's disease.

4. Discussion and Future Work

The Patent EnrichMent Tool (PEMT) is an open-source and easy-to-use tool providing the scientific community with the opportunity to extract pharmaceutically relevant patent documents for genes of interest. This efficient linking of genes to patents could potentially open a link to enrich existing biomedical knowledge graphs with the unmined patent literature. Moreover, PEMT can help in providing an overview of the target prioritization shift over the time period. Hence, with PEMT we hope to enhance the accessibility and applicability of patent literature within the scientific community.

Funding

Conflict of Interest: none declared.

References

Chae, Y. K., et al. (2017). Inhibition of the fibroblast growth factor receptor (FGFR) pathway: the current landscape and barriers to clinical application. Oncotarget, 8(9), 16052.

Donald, K. E. et al. (2018). Tips for reading patents: a concise introduction for scientists. Expert Opinion on Therapeutic Patents, 28(4), 277-280.

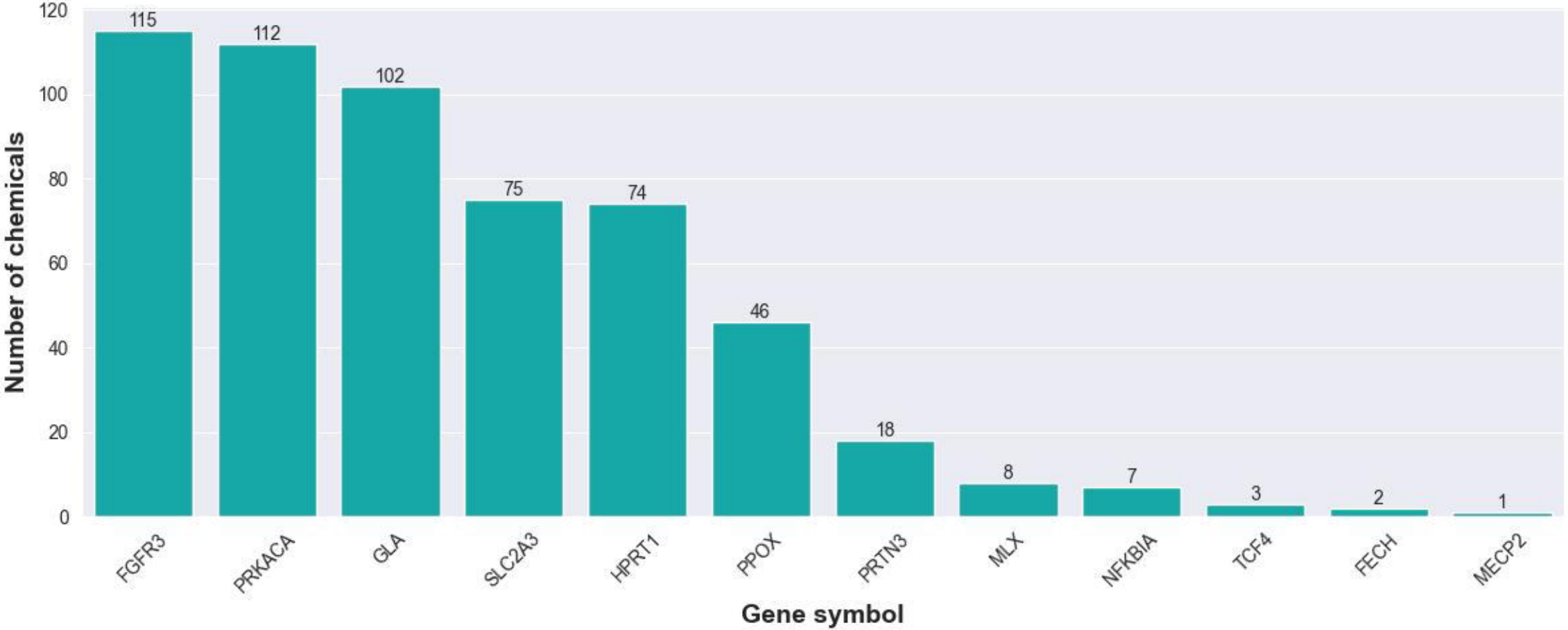
Grand, E. K., et al. (2004). Targeting FGFR3 in multiple myeloma: inhibition of t (4; 14)-positive cells by SU5402 and PD173074. Leukemia, 18(5), 962-966.

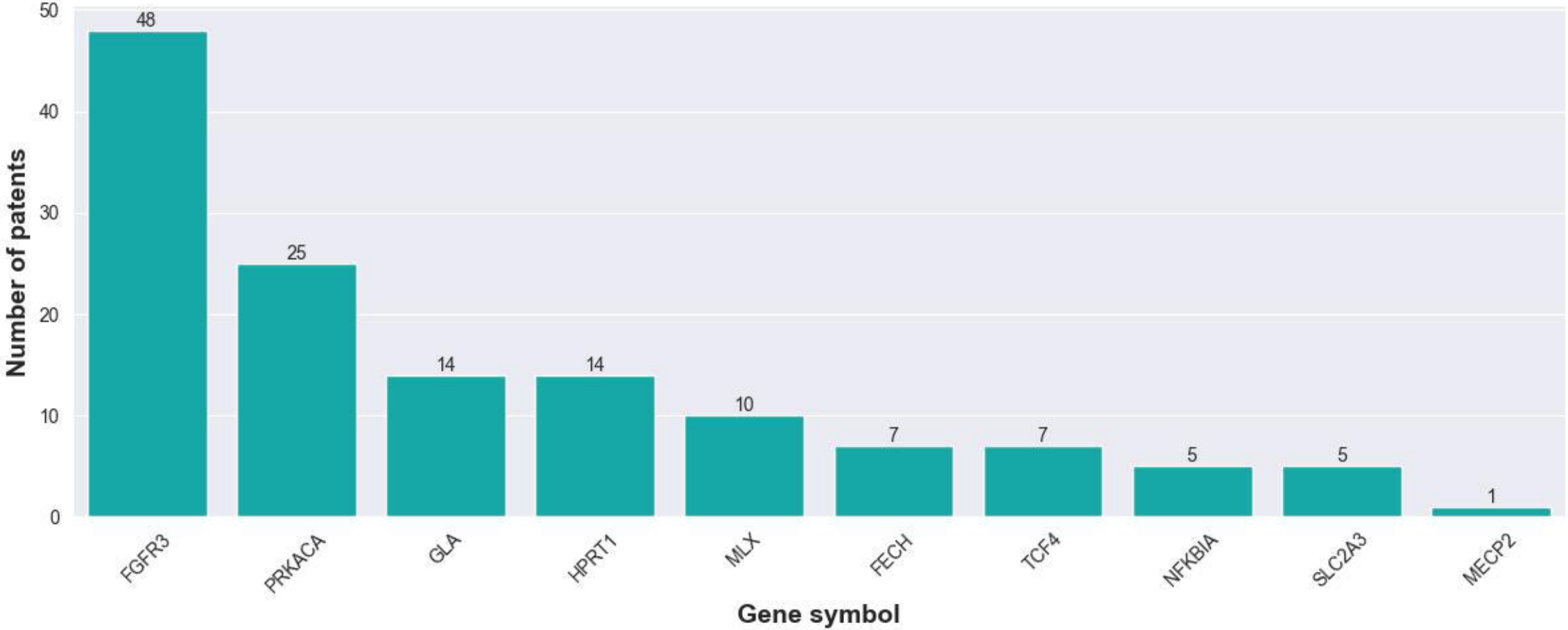
Jin, G., and Wong, S. T. (2014). Toward better drug repositioning: prioritizing and integrating existing methods into efficient pipelines. Drug discovery today, 19(5), 637-644.

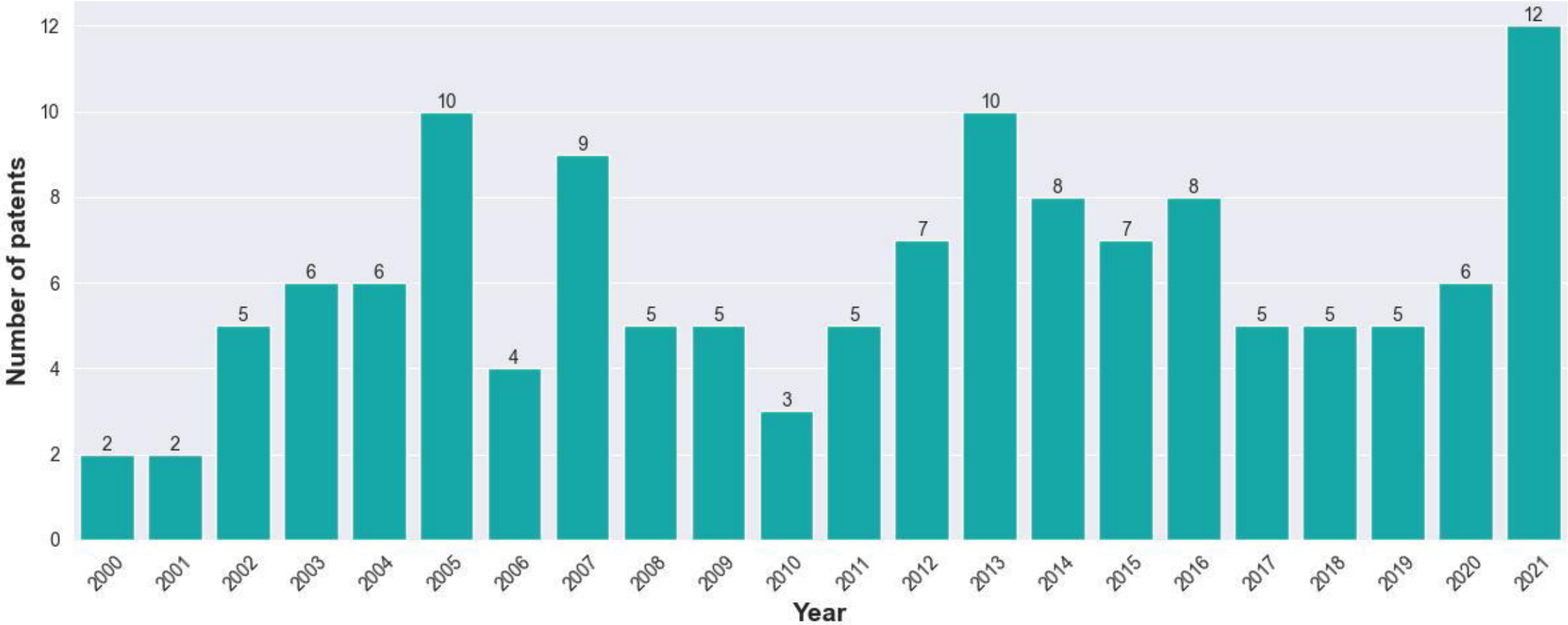
Mucke, H. A. (2021). What patents tell us about drug repurposing for cancer: A landscape analysis. In Seminars in Cancer Biology, 68, 3-7

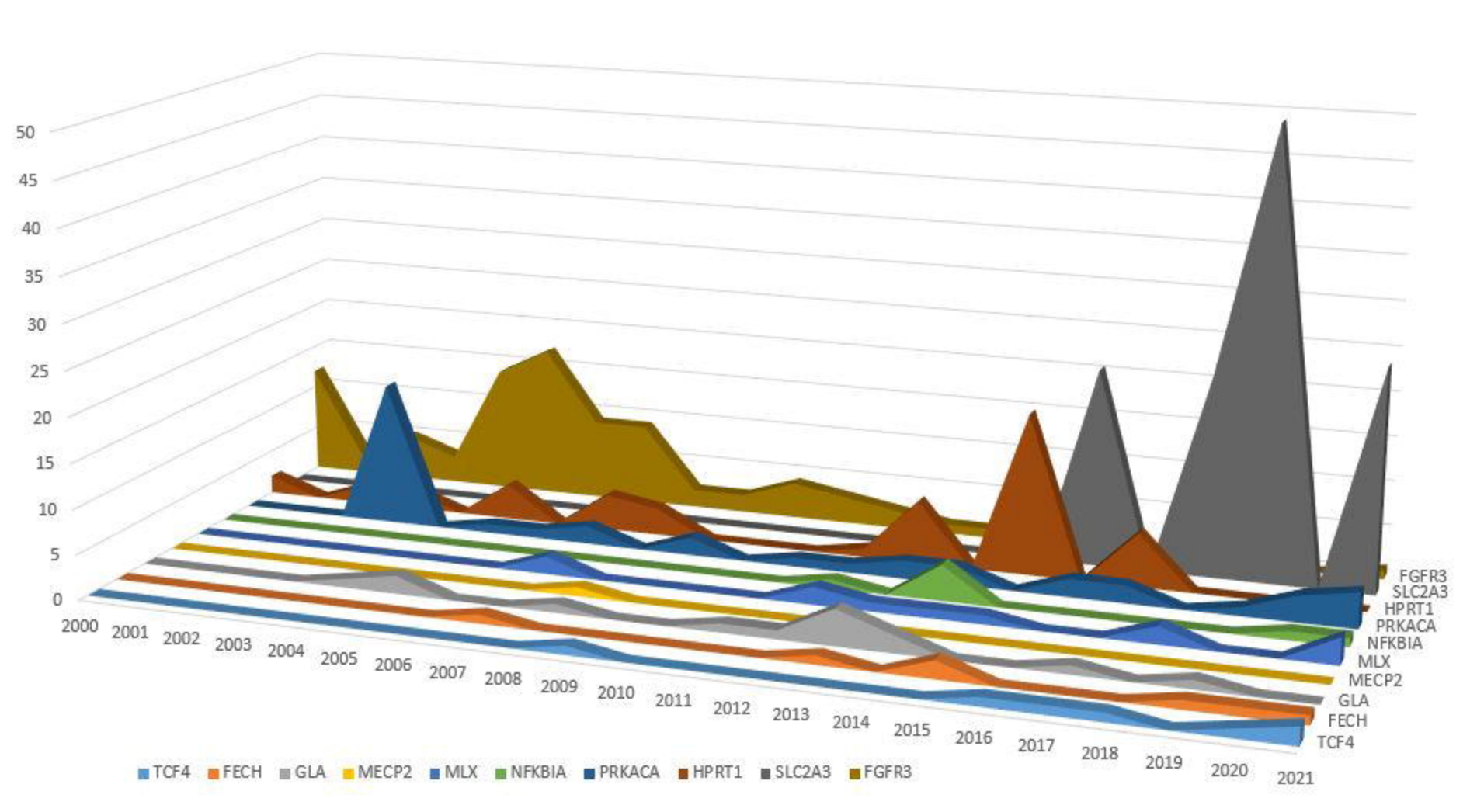
Roskams-Edris, D., et al (2019). Medical Methods Patents in Neuromodulation. Neuromodulation: Technology at the Neural Interface, 22(4), 398-402.

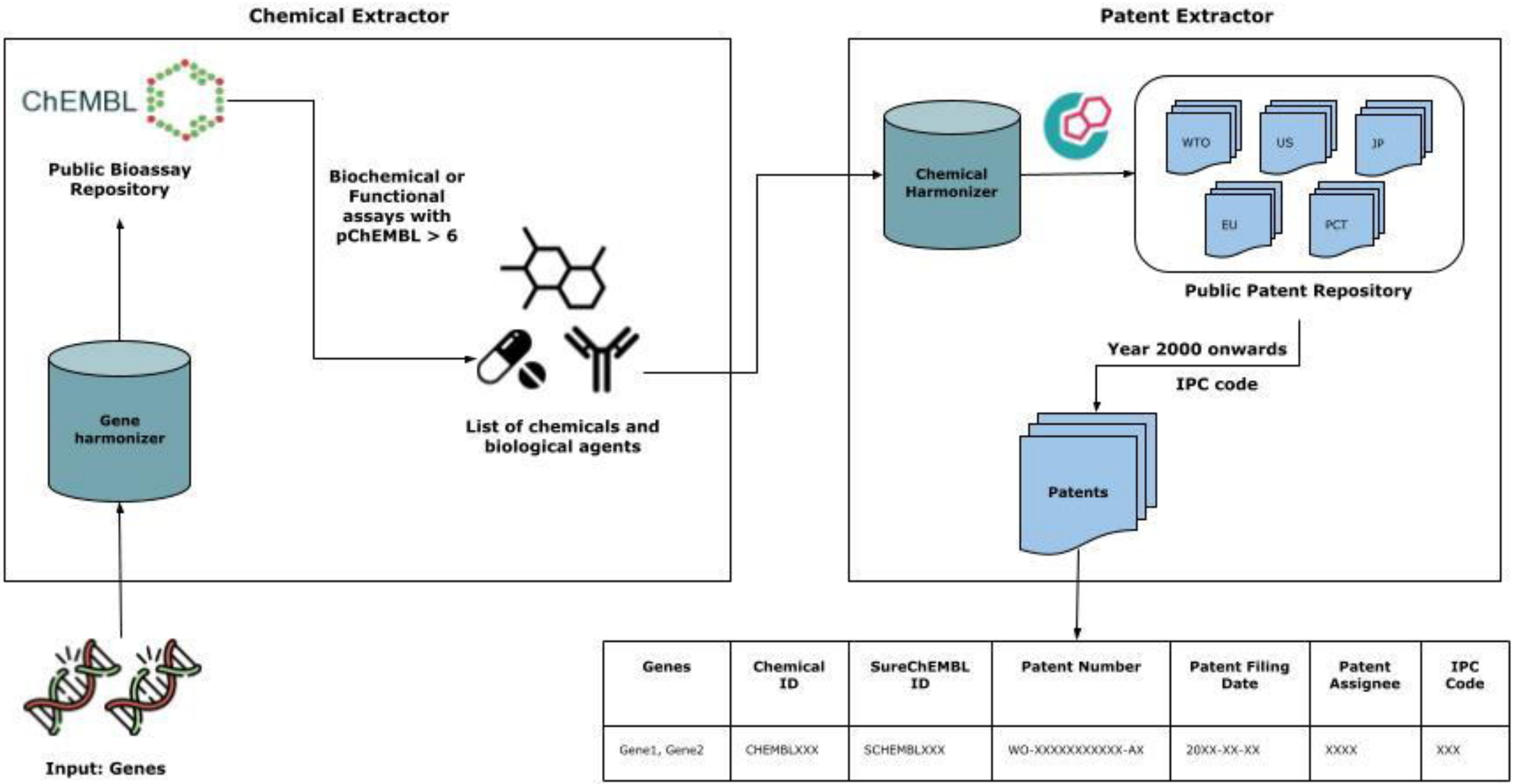
Weinreich, S. S., et al (2008). Orphanet: a European database for rare diseases. Nederlands tijdschrift voor geneeskunde, 152(9), 518-519.


Number of chemicals
120
100
80
60
40
20
0
115
112
102
75
74
46
18
8
7
3
2
1
FGFR3
PRKACA
GLA
SLC2A3
HPRT1
PPOX
PRTN3
MLX
NFKBIA
TCF4
FECH
MECP2
Gene symbol


Number of patents
50
40
30
20
10
0
48
25
14
14
10
7
7
5
5
1
FGFR3
PRKACA
GLA
HPRT1
MLX
FECH
TCF4
NFKBIA
SLC2A3
MECP2
Gene symbol


Number of patents
0
2
4
6
8
10
12
2
2
5
6
6
10
4
9
5
5
3
5
7
10
8
7
8
5
5
5
6
12
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
Year


50
45
40
35
30
25
20
15
10
5
0
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
FGFR3
SLC2A3
HPRT1
PRKACA
NFKBIA
MLX
MECP2
GLA
FECH
TCF4
TCF4
FECH
GLA
MECP2
MLX
NFKBIA
PRKACA
HPRT1
SLC2A3
FGFR3


Chemical Extractor
ChEMBL
Public Bioassay Repository
Biochemical or Functional assays with pChEMBL > 6
Gene harmonizer
List of chemicals and biological agents
Input: Genes
Patent Extractor
Chemical Harmonizer
WTO
US
JP
EU
PCT
Public Patent Repository
Year 2000 onwards
IPC code
Patents
Genes
Chemical ID
SureChEMBL ID
Patent Number
Patent Filing Date
Patent Assignee
IPC Code
Gene1, Gene2
CHEMBLXXX
SCHEMBLXXX
WO-XXXXXXXXXXX-AX
20XX-XX-XX
XXXX
XXX
Output: Gene linked patents