universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Analysis of patent bioactivity data in ChEMBL"

verfasst von / submitted by

Ursula Bartuschka BSc MSc

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2023 / Vienna 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 606
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Drug Discovery and Development
Betreut von / Supervisor:	Mag. Dr. Barbara Zdrazil, Privatdoz.

Acknowledgements

When I started this master's program, it was important to me to learn something completely new, but I did not yet know what form that would take. The two courses I had with Barbara Zdrazil awakened my interest in data science and cheminformatics, and I am extremely glad that she agreed to be my thesis supervisor. Her support encouraged me to grow my knowledge, sometimes in unexpected directions, and her comments and suggestions helped to motivate me whenever I felt stuck. I am also grateful for her continuing the supervision even through her change of employment from the University of Vienna to EMBL-EBI. I could not have wished for a more harmonious and fruitful collaboration.

Pursuing another degree would have been impossible without many people around me. First and foremost, my husband Alexander Bartuschka, who has been an invaluable source of support and inspiration throughout this process, and who I am incredibly lucky to share my life with. Another source of unwavering encouragement has been my grandmother, Ilse Gebauer, whom I want to thank sincerely for always believing in me. My best friend Claudia Hüttmann was sympathetic and uplifting whenever we met, even though I had so little time during the last years.

Many thanks are due to all my friends and colleagues, who always were there when I needed to talk. Switching from full-time to part-time work required flexibility and understanding especially by Sandra Jost, who always put me in a better mood when working together and never once complained about the extra work (I'm sorry!). I am also very grateful to my previous manager Urban Lundberg, who backed my decision to study again wholeheartedly.

Abstract

In drug discovery and development, there is a tendency to focus on a small subset of targets and target families, while leaving a considerable part of the human proteome relatively unexplored. A lack of knowledge for these understudied targets leaves them unattractive for further research in a self-reinforcing cycle.

One previously largely untapped potential source of data on understudied targets are patents, which are both numerous and notoriously hard to extract bioactivities from. Nevertheless, in recent years efforts have been made to incorporate such data into the ChEMBL database to facilitate the study of previously underexplored targets.

This thesis aimed to evaluate the gain in knowledge from ChEMBL's patent annotation efforts. Each target's patent compounds are analyzed from two perspectives: Firstly, regarding their diversity, giving an indication of the coverage of each target, and secondly regarding their novelty when compared to compounds from other sources for the same target.

Going forward, such analyses may help to guide the prioritization of targets for patent annotations towards targets which are both in need of more data, and where patents hold great potential for novel compounds.

Zusammenfassung

Die Forschung und Entwicklung von neuen pharmazeutischen Wirkstoffen konzentriert sich meist auf nur wenige unterschiedliche Targets und Targetfamilien, während ein großer Teil des menschlichen Proteoms vergleichsweise wenig erforscht wird. Der Mangel an Daten für jene Targets führt wiederum dazu, dass sie weiterhin wenig attraktiv für die Forschung bleiben.

Eine bisher wenig genutzte Datenquelle für solche wenig erforschten Targets sind Patente. Deren Anzahl und Struktur erschweren es, Bioaktivitäten für die enthaltenen Moleküle zu extrahieren. Trotzdem wurden in den letzten Jahren Anstrengungen unternommen, solche Daten in die ChEMBL Datenbank zu integrieren, und die Erforschung bisher vernachlässigter Targets zu erleichtern.

Das Ziel dieser Arbeit war, den Gewinn an Informationen durch die Annotation von Patenten durch ChEMBL zu evaluieren. Für jedes Target wurden die Moleküle aus Patenten in zweierlei Hinsicht analysiert: Erstens, in Bezug auf ihre Diversität, was Schlüsse auf die Abdeckung des Targets aus Patenten in ChEMBL erlaubt, und zweitens hinsichtlich ihrer Neuartigkeit im Vergleich mit Molekülen für das gleiche Target aus anderen Datenquellen.

Diese Analysen können einen Beitrag dazu leisten, Targets für die Annotation aus Patenten zu priorisieren, sodass Daten bevorzugt für jene extrahiert werden, für die noch wenig Informationen verfügbar sind, aber Patente großes Potential für neuartige bioaktive Moleküle bergen.

Table of contents

List of figures

List of tables

List of abbreviations

ADMET	Absorption, Distribution, Metabolism, Excretion, Toxicity
CDP	Consensus Diversity Plot
EPO	European Patent Office
FDA	Food and Drug Administration
GPCR	G-protein Coupled Receptor
IDG	Illuminating the Druggable Genome
InChI	International Chemical Identifier
JPO	Japan Patent Office
MACCS keys	Molecular Access System keys
MoA	Mechanism of Action
NIH	National Institutes of Health
NR	Nuclear Receptor
oGPCR	olfactory G-protein coupled receptor
TDL	Target Development Level
TF	Transcription Factor
SAR	Structure-Activity Relationship
SMILES	Simplified Molecular Input Line Entry System
USPTO	United States Patent and Trademark Office
WIPO	World Intellectual Property Organization

1 Introduction

1.1 Understudied targets in drug discovery and development

A central paradigm in pharmaceutical research is that drugs exert their function by interacting with a target, which results in a disease-modifying effect. A target in this context can be broadly defined as a biological entity – such as a protein, gene or RNA – which a compound interacts with, often modulating the entity's function, or disrupting interactions with other partners[1]. A good drug target does not only produce the desired effect but provides the opportunity to do so efficaciously and safely, with as few side or off-target effects as possible. The identification of a potentially suitable drug target, thus, is not sufficient to initiate a research program. In the next step of target validation a body of knowledge has to be established around the target, covering its biological significance and pathways it is involved in[1,2].

As this process is time-consuming and work-intensive, the decision to investigate a certain target is not taken lightly. The selection of drug targets for pharmaceutical research, which is to a large part carried out by commercial enterprises, is strongly influenced by strategic and financial considerations in addition to experimental evidence[3,4]. To minimize risk in drug development, research tends to focus on a relatively small subset of targets which are well-explored[5–7], while other targets remain understudied. Possible reasons why a target is not explored further include a lack of tools for study (such as chemical probes or monoclonal antibodies), insufficient funding opportunities for research of riskier targets, a lack of scientific interest, or a detrimental combination of all of these influences[6].

Around seventy percent of FDA-approved small molecule drug target proteins from only four families: GPCRs, ion channels, kinases and nuclear receptors[8]. While many drugs predate the sequencing of the human genome, the predominance of these families has not been altered much by the possibilities of the post-genomic era, probably at least in part due to the long development times of drugs[8,9]. Some targets are also harder to develop compounds for than others: For example, proteins without enzymatic activity frequently have posed more of a challenge for finding small molecule ligands modulating their effects[10], possibly explaining their relative scarcity among current drug targets.

The Illuminating the Druggable Genome (IDG) initiative, started in 2014 by the US National Institutes of Health (NIH), seeks to map available information on human drug targets to identify knowledge gaps for potentially druggable proteins that have so far been neglected, and induce further research on them[5]. The initial focus on the subset of highly druggable target families has since been expanded to include the entire human proteome[11].

To classify targets by the amount of data available for them, the IDG developed the Target Development Level (TDL) classification system[5]. The system consists of four levels: Tclin, Tchem, Tbio and Tdark; the breakdown of the proteome into these levels and further into target families (as of 2018, but changes since then have been minor) is shown in Figure 1.

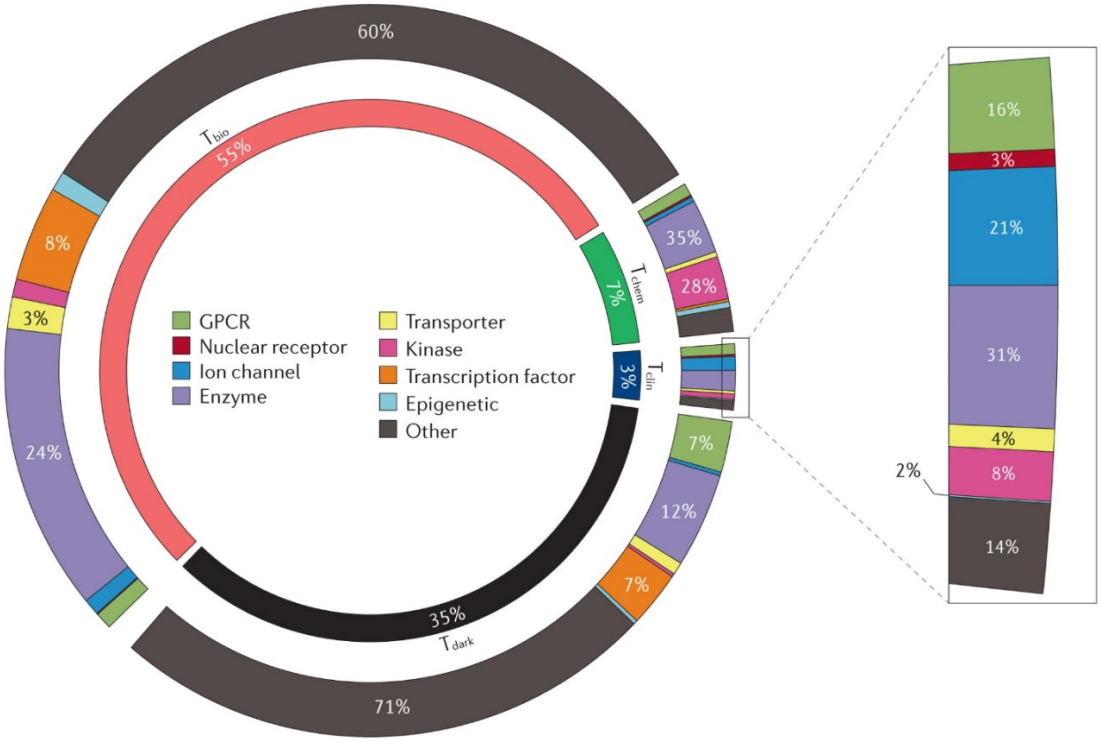


Figure 1: Target Development Levels and numbers by target families. Adapted from Figure 1a in ref. 5. *The inner ring represents the percentages of the four TDLs of the whole proteome, which are subdivided into families for each TDL in the outer ring. While some reclassifications have taken place since the publication of this figure in 2018, there have been no major changes to the overall picture.*

In brief, Tclin targets are differentiated from the other classes by the availability of at least one approved drug linked to it by a mechanism of action (MoA). Tchem targets do not have an associated drug, but at least one active compound, binding with a potency above a defined cutoff value. This value differs by protein family and was chosen based on the bioactivity of approved drugs binding targets of that family (which can be found in the methods section in Table 5). Currently, only 704 proteins are in the Tclin category, representing 3.4% of targets catalogued by the IDG initiative. It is important to note that unambiguous assignment of an MoA is not always straightforward[8] and therefore many targets interacting with approved drugs are at present in the Tchem category[5].

Targets in the two other levels, Tbio and Tdark, are commonly regarded as understudied and in need of further exploration[5] . Tbio targets lack active ligands conforming to the defined cutoff values, or do not meet the MoA criterion required for a Tclin classification, while still meeting other conditions showing their relevance: a Mendelian disease phenotype in the OMIM database (Online

Mendelian Inheritance of Man)[12], Gene Ontology leaf annotations from experimental evidence, or two out of three other criteria related to prevalence in literature, functional information, and commercial antibody availability. Tdark comprises targets which are manually reviewed in UniProt, but while other data about them may be available, they do not fulfill the other levels' criteria.

However, not every human protein has the inherent potential to be a therapeutically relevant drug target, as those are defined by being both disease-modifying and druggable[9,13], and some proteins may forever remain in Tdark; an estimate from 2002 was for 600-1,500 druggable small molecule targets in humans[9]. The data collected and processed by the IDG initiative resources can be accessed and downloaded via the web interface Pharos[14]. When more data becomes available upon new releases from the source databases, TDLs are recalculated and updated[11].

Among the most important tools for studying the function of a potential drug target are small molecule modulators[15] which interact specifically with the target (chemical probes). A wider availability of data on such chemical probes is expected to aid in illuminating understudied targets[6].

1.2 Patents as a source of bioactivity data

One potentially very valuable source of bioactivity data for the study of less explored targets, but at the same time one that is hard to make full use of due to various factors, are patents. It has been argued that information from patents might in some respects be more directly applicable than from scientific literature, as the cost and effort of patenting may only be worthwhile if researchers have a certain anticipation of a patent's future value[16]. However, several studies have found that data from pharmaceutical patents may not become available through scientific literature until several years later, if at all[16–18].

While patent applications are publicly available through patent authorities and search engines, they are not particularly accessible in that form, especially in the domain of chemistry. Although much progress has been made in identifying useful patents and extracting relevant information, it remains a nontrivial task[17,19]. Automated approaches such as text-mining have required specialized strategies to be developed to cope with the structural and stylistic peculiarities of patents that differ from scientific literature, but also to correctly identify compounds by chemical names, extract structures from images, and interpreting Markush structures[16].

Also, as patents may contain hundreds of molecules, not all of which are equally important, techniques have been developed to identify the key compounds[20,21], or to distinguish relevant compounds from common reagents, starting materials and intermediate products[22]. None of these tasks is easily implemented or error-free, and manual extraction may be preferred for reasons of

reliability over automated processes despite the slower turnaround time and high labor requirement[18,23].

Previously the chemical annotation and curation of relevant patents was largely accessible only through commercial data collections, but in more recent years, large-scale open databases such as SureChEMBL have made chemistry-related patent information much more accessible to the public[22,24]. The SureChEMBL database is continuously updated with data from full-text patents from the European Patent Office (EPO), the United States Patent and Trademark Office (USPTO), and the World Intellectual Property Office (WIPO), as well as abstracts from the Japanese Patent Office (JPO). Besides bibliographic information, SureChEMBL automatically extracts compounds from patent text, images and molfile attachments[24]. Dealing with the sheer amount of – to date, over 25 million compounds have been extracted from an approximately equal number of chemistry-containing patents by SureChEMBL, with a monthly addition of 80,000 more compounds[17] – clearly requires automated processes to extract and process the data into an accessible form[24].

SureChEMBL provides detailed functionality for patent searching and filtering, including substructure searches in the patent corpus. This is crucial for identification of patents that may contain measured compound bioactivities on a target of interest[17]; however, it is currently not equipped to extract such relationships. This is a complicated task that requires expert knowledge and experience and therefore manual annotation to avoid errors.

1.3 Bioactivities in ChEMBL

Launched over a decade ago, with the first release made available in 2009, ChEMBL[25] is an open-access bioactivity database of considerable size, comprising around 2.4 million compounds tested in approximately 1.5 million assays against 15,000 targets in its latest release at the time of writing (ChEMBL 32). Bioactivity data is manually extracted from seven core medicinal chemistry journals as well as from selected articles from other journals; in addition, data is exchanged with other databases, and datasets may be directly deposited[26]. To date, the share of bioactivities in ChEMBL from deposited data sets vs. data extracted from literature is roughly 1:1.

After the initial extraction of bioactivities from any of these sources follows a curation and standardization process that corrects common errors and ensures proper incorporation of the data[27]. One of the most important and complex steps is the assignment of a target. UniProt identifiers are used for protein targets, but ChEMBL captures targets in more detail: here, they are defined as “the entity with which the compound actually interacts in a particular assay system”[27], which in many cases is not a single protein, but may (among other possibilities) be a complex with other proteins, a protein-protein interaction, or a non-protein entity such as DNA.

To distinguish between targets and their molecular components, different target types are defined. In cases where assays are carried out in cells, tissues or whole organisms, it may not be possible to attribute the effect to a smaller entity. Similarly, for assays with a group of very closely related proteins as targets, where one cannot be sure of the exact receptor subtypes or protein isoform responsible for an effect, the assigned target type is "protein family", as it would be misleading to map it to individual members; the same type is used where there is nonselective binding to all members of a family. A comparable approach is used for protein complex targets. Further ambiguity is introduced when both of these conditions are met: a protein complex where the exact composition is not known; the target type is then given as "protein complex group"[28]. For these reasons, multiple target types may be associated with the same UniProt identifier; they are distinguished in ChEMBL by assigning each a different target ID. The most common distinct target types are depicted with examples in Figure 2.

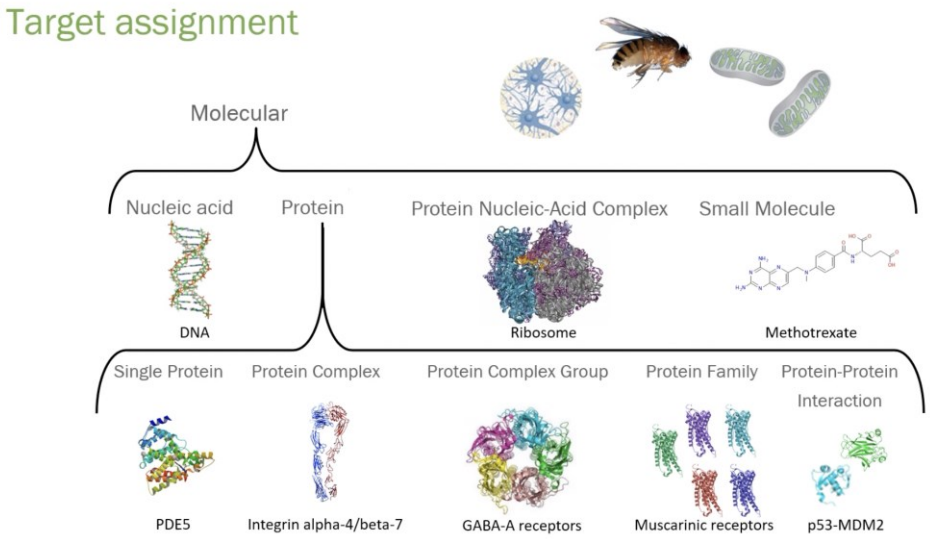


Figure 2: Target types in ChEMBL. Examples for target types commonly found in ChEMBL are shown. Non-molecular targets include cells and tissues, whole organisms and organelles. This figure was kindly provided by the ChEMBL team.

In order to expand the bioactivity data included in the ChEMBL database beyond what is available from sources such as articles in scientific journals and donated data sets, in recent years activity measurements from patents have been incorporated as well, either through manual extraction by ChEMBL or exchange of data with BindingDB[17,29,30]. However, the number of patents potentially worth annotating in the world of chemistry and drug discovery is vast. Any effort for manual extraction must be limited to specific topics or targets of interest, and an exhaustive or complete annotation may be out of reach for any but a very small research area at this point in time.

ChEMBL has decided to focus their patent annotation efforts on understudied targets, and details on the process of identifying which patents are likely to contain bioactivity data for understudied (Tdark or Tbio) targets in order to prioritize them for manual annotation were published recently[17].

1.4 Aims of this project

This work aims to examine the information that has so far been added to ChEMBL from patent annotations and to gauge if and how it has complemented the data the database holds from curation of data from other resources. An attempt is made to estimate the value added by curating compound bioactivities from patents, which may lie in extracting bioactivities for targets that have no annotations from non-patent literature in ChEMBL yet, or in the expansion of chemical space by annotating bioactivities from a more diverse set of compounds for a target, especially concerning new active compounds. This information may then be used to guide prioritization efforts towards those targets that are most in need of (further) illumination.

To accomplish this, a target-centered perspective is used. For each human target where patent bioactivities have been added to ChEMBL, the compounds from patents are analyzed in terms of their diversity, and in the second part, their novelty as compared to compounds annotated for the same target from non-patent sources.

There are different ways to estimate the chemical diversity (or its inverse, similarity) of a dataset, usually a chemical library. All of them have different uses and come with advantages and disadvantages. Scaffold diversity, for example, may be used for the characterization of screening libraries to avoid a structurally too homogenous dataset[31]. It is easily interpretable, but by its nature omits any more detailed structures in the side chains of the compounds analyzed[32]. Compound diversity may be assessed by different bit-string fingerprints depending on their performance for a desired application, combined with a suitable similarity coefficients[33,34]. Their common drawback is that they are not easily interpretable. A third possibility is to estimate the physicochemical diversity of a collection of molecules along several properties known to play a role in drug-likeness; but very different molecules may share a similar physicochemical profile, aside from its suitability for the task[32].

Using a combination of available diversity metrics may be preferable to avoid the disadvantages of any one of them. This makes it possible to analyze the diversity metrics not independently from each other, but to integrate their results and, in addition, visualize them in an intuitive manner as a "consensus diversity plot" (CDP)[32]. While the article describing this tool used whole databases as subjects to compare, it is utilized here to analyze the diversity (or novelty) of individual target's patent compounds. Considering that no specific novelty aspect is favored at this stage, it would be less appropriate to assess diversity of the patent bioactivity annotations with only one of the possible methods.

The novelty of patent compounds is defined in this project as the inter-dataset dissimilarity between compounds from patents and those from non-patent sources for the same target in ChEMBL (in

contrast to the intra-set dissimilarity for one target's patent compound diversity). While the name may be misleading - no temporal component is involved - it refers to the novel chemical space made accessible through patent annotations for a target where compounds are added from patents that are quite different from those already published elsewhere.

Importantly, it would be wrong to assume that for every target the chemical space that could potentially be uncovered through bioactivity annotations, from patents or other sources, is equally large. Some targets may not be tractable in terms of facilitating the binding of compounds (even with low affinity) outside a small range of scaffolds, side chains or physicochemical parameters. Even though considering the number of patent annotations currently available in ChEMBL such a scenario appears unlikely, it may still occur for non-patent bioactivities, which would result in low novelty.

Another point to keep in mind is that while ChEMBL holds a considerable amount of data from non-patent sources and continuously extracts articles with relevant medicinal chemistry data, the patent annotations are far from exhaustive. For any of the targets analyzed in this work, many more patents may exist that simply have not been identified or annotated yet. Sparse patent data in ChEMBL for a particular target thus does not indicate that such data does not exist; it may simply not have been identified or extracted, or has not been prioritized.

2 Methods

This section describes how the workflow for extracting the data from ChEMBL, processing and splitting of the datasets, and calculation of the similarity metrics was developed and refined over the course of this thesis, and what influenced the decisions made therein.

2.1 Software

All work was carried out using Python version 3.9.5 in Jupyter Notebooks (version 6.4.0). The libraries used were pandas (version 1.3.4), numpy (1.20.1), rdkit (2021.03.5), scikit-learn (1.0.1), scipy (1.7.1), and all diagrams were made using seaborn (0.11.2) and/or matplotlib (3.4.3). For querying the ChEMBL database, sqlite3 and the ChEMBL downloader (https://github.com/cthoyt/chembl-downloader) were used.

2.2 Extraction of data on understudied targets from ChEMBL

For assessing the coverage of understudied (Tdark or Tbio) targets in ChEMBL 31 of the database (which was the current release at the time the work was initiated) was queried using the ChEMBL downloader with a list of target accession codes (5,679 for Tdark and 12,058 for Tbio targets) obtained from the Pharos web interface[11] (https://pharos.nih.gov). The ChEMBL database schema can be found in the appendix (section 6.1).

Fields pertaining to compound bioactivities were extracted per the following query:

```
SELECT
component_sequences.accession,
target_dictionary.tid,
target_dictionary.target_type,
assays.doc_id,
assays.assay_type,
docs.doc_type,
docs.year,
activities.pchembl_value,
activities.molregno,
molecule_dictionary.chembl_id,
compound_structures.standard_inchi,
compound_structures.canonical_smiles
FROM component_sequences
LEFT JOIN target_components ON component_sequences.component_id =
target_components.component_id
LEFT JOIN target_dictionary ON target_components.tid = target_dictionary.tid
LEFT JOIN assays ON assays.tid = target_dictionary.tid
LEFT JOIN docs ON docs.doc_id = assays.doc_id
LEFT JOIN activities ON activities.assay_id = assays.assay_id
LEFT JOIN molecule_dictionary ON molecule_dictionary.molregno =
activities.molregno
LEFT JOIN compound_structures ON compound_structures.molregno =
activities.molregno
WHERE component_sequences.accession IN {tuple(targetlist)}
```

Due to the way this query was structured, also some entries without associated "MOLREGNO" were returned, which contain only some information on the target while lacking bioactivities. These entries (18 for Tdark targets, 366 for Tbio targets) were dropped before further analyses.

2.3 Extraction of patent bioactivity data from ChEMBL

The initial query for patent bioactivities in ChEMBL 31 was made via the ChEMBL downloader. As it was at that point not yet clear what direction the analyses would take, a relatively expansive number of fields, not all of which were relevant to the final workflow, were selected via left join operations of the respective tables. The full query was as follows:

```
SELECT
docs.doc_id,
docs.patent_id,
docs.year,
docs.ridx,
docs.src_id,
assays.tid,
assays.doc_id as assay_doc_id,
assays.assay_type,
target_dictionary.target_type,
target_dictionary.pref_name,
target_dictionary.tax_id,
target_dictionary.organism,
target_dictionary.chembl_id as target_chembl_id,
component_sequences.accession,
protein_classification.pref_name,
protein_classification.short_name,
protein_classification.protein_class_desc,
protein_classification.definition,
activities.molregno,
activities.pchembl_value,
compound_records.compound_name,
molecule_dictionary.chembl_id,
molecule_dictionary.max_phase,
compound_structures.standard_inchi,
compound_structures.canonical_smiles
FROM docs
LEFT JOIN assays ON assays.doc_id = docs.doc_id
LEFT JOIN target_dictionary ON target_dictionary.tid = assays.tid
LEFT JOIN target_components ON target_components.tid = target_dictionary.tid
LEFT JOIN component_sequences ON component_sequences.component_id =
target_components.component_id
LEFT JOIN component_class ON component_class.component_id =
target_components.component_id
LEFT JOIN protein_classification ON protein_classification.protein_class_id =
component_class.protein_class_id
LEFT JOIN activities ON activities.assay_id = assays.assay_id AND
activities.doc_id = assays.doc_id
LEFT JOIN compound_records ON compound_records.molregno = activities.molregno
LEFT JOIN molecule_dictionary ON molecule_dictionary.molregno =
compound_records.molregno
LEFT JOIN compound_structures ON compound_structures.molregno =
compound_records.molregno
WHERE doc_type IN ('PATENT')
```

Of note, this query includes both sources of patent data in ChEMBL 31, which are ChEMBL's own patent annotations (source ID 38), and data obtained by exchange with BindingDB (source ID 37).

2.4 Processing of the patent query results

2.4.1 General considerations and scaffold generation

The dataset retrieved by the query described in section 2.3 consisted of 1,538,164 entries from 2,439 patents, corresponding to 178,939 unique compounds (Table 1).

Table 1: Unique values retrieved for selected fields.

Field retrieved	Number of unique values
doc_id	2,447
patent_id	2,439
year	16
src_id	2
target_id	2,180
assay_doc_id	2,447
target_organism	101
target_chembl_id	2,180
accession	1,946
molregno	178,939
cmpd_chembl_id	178,939
smiles	178,595

Of note, the number of DOC_IDs is slightly larger than that of PATENT_IDs. This is due to eight patents in the dataset possessing two different DOC_IDs for unclear reasons (Table 2). To avoid errors resulting from this, only PATENT_ID is used for any operations later on.

Table 2: Patents with two different DOC_IDs.

PATENT_ID	DOC_IDs
US-8536198-B2	94121, 113801
US-8609708-B2	94234, 113911
US-20140128392-A1	97144, 113785
US-20150031679-A1	97146, 113833
US-20160237082-A1	109392, 113906
US-20160304533-A1	109355, 113880
US-8466108-B2	102346, 113865
US-20120225846-A1	97143, 113781

At this point, Bemis-Murcko scaffolds were generated using the rdkit.Chem.scaffolds function MurckoScaffold. While this was successful for most compounds, 344 compounds did not have a SMILES deposited in ChEMBL and therefore no scaffold could be generated. A manual spot check revealed that this appears due to these compounds possessing coordinated metal bonds, which cannot be accurately represented by standard InChI and for which no structure is provided by ChEMBL for that reason (as described in the ChEMBL documentation: https://chembl.gitbook.io/chembl-interface-documentation/frequently-asked-questions/drug-and-

compound-questions#can-you-provide-more-details-on-the-removal-of-metal-containing-compounds). Such compounds were removed from the dataset in a later step.

2.4.2 Target type and assay type filtering

The targets in the dataset had a variety of target types (Table 3). As the focus of this study was on protein targets, a decision was made to only include targets of the type "single protein" and "protein complex". This also excludes the "protein complex group" type, which comprises protein complexes of unclear composition, and may give misleading attributions to understudied targets where their involvement is not confirmed.

Table 3: Target types in the initial dataset.

Target type	Number of targets
SINGLE PROTEIN	1,701
CELL-LINE	256
PROTEIN COMPLEX	96
PROTEIN FAMILY	45
ORGANISM	35
TISSUE	16
PROTEIN COMPLEX GROUP	15
PROTEIN-PROTEIN INTERACTION	7
CHIMERIC PROTEIN	2
UNCHECKED	1
NON-MOLECULAR	1
ADMET	1
UNKNOWN	1
NO TARGET	1
NUCLEIC-ACID	1
SUBCELLULAR	1

As the focus of this project is on specific target bioactivities, the assay types were also restricted to binding (B) and functional (F) assays, excluding ADMET and toxicity assays as well as uncategorized assays.

2.4.3 Target organism restriction

The processing steps described so far yielded a dataset consisting of 713,215 entries for 1,786 targets (counted as unique UniProt identifiers, not ChEMBL's target IDs), the vast majority of which are human targets (Table 4). Other organisms with substantial numbers of patent bioactivities are model organisms such as mice or rats. A range of targets from viruses and other pathogens (bacterial, protozoic or fungal) is among the extracted targets as well. While it would certainly be interesting to look more closely into patent bioactivities for other target organisms, especially pathogens, there is no systematic information on how well such targets are studied so far, so an extensive literature review would be required to make such a classification. An additional difficulty

is that for example for bacteria and viruses, proteins may differ by strain in important ways and thus bioactivities may only be valid for the specific strain tested. It is therefore hard to define the level of exploration for such targets.

Table 4: Number of targets for the most numerous target organisms in the dataset. Note that the entry for Escherichia coli combines the numbers for the two versions "Escherichia coli (strain K12)" and "Escherichia coli K-12", but excludes those for "Escherichia coli" alone, as no strain is given.

Target Organism	Number of targets
Homo sapiens	1,370
Mus musculus	139
Rattus norvegicus	126
Bos taurus	19
Escherichia coli (strain K12)	12
Staphylococcus aureus	8
Sus scrofa	8
Oryctolagus cuniculus	7
Cricetulus griseus	6
Plasmodium falciparum	5
Hepatitis C virus	5

As detailed in the introduction, the focus of patent annotation efforts by ChEMBL is on understudied targets with a Tdark or Tbio designation by the IDG initiative. The scope of this project was therefore defined to exclusively include human targets. After excluding all other target organisms, 652,277 entries (comprising 158,601 unique compounds) with bioactivities for 1,370 human targets remained in the dataset, extracted from 2,172 patents.

2.5 Query for non-patent literature annotations

For the set of 1,370 targets fulfilling the criteria outlined in the previous sections, another query to ChEMBL was constructed to extract all associated bioactivities originating from *non-patent* literature:

```
SELECT
target_dictionary.tid,
target_dictionary.target_type,
component_sequences.accession,
assays.doc_id,
assays.assay_type,
docs.doc_type,
docs.year,
activities.pchembl_value,
activities.molregno,
molecule_dictionary.chembl_id,
compound_structures.standard_inchi,
compound_structures.canonical_smiles
FROM target_dictionary
LEFT JOIN assays ON assays.tid = target_dictionary.tid
LEFT JOIN target_components ON target_components.tid = target_dictionary.tid
LEFT JOIN component_sequences ON component_sequences.component_id =
target_components.component_id
LEFT JOIN docs ON docs.doc_id = assays.doc_id
LEFT JOIN activities ON activities.assay_id = assays.assay_id
```

```
LEFT JOIN molecule_dictionary ON molecule_dictionary.molregno =
activities.molregno
LEFT JOIN compound_structures ON compound_structures.molregno =
activities.molregno
WHERE target_dictionary.tid IN {tuple(targets_patents_human)} AND
docs.doc_type != 'PATENT'
```

2.5.1 Processing of non-patent data query results

The dataset retrieved from the query described above consisted of 855,566 unique compounds for 1,257 targets from 29,887 source documents. The vast majority of entries was from deposited datasets (1,866,585) and publications (1,738,113), but 373 entries were also added from books. Most of the processing steps for the patent data were not necessary to repeat for the non-patent data, as the latter query specified target IDs that were already the correct target organism and target type. Thus, the only necessary steps were to exclude all assay types except “B” and ”F” and generate Bemis-Murcko Scaffolds for the remaining compounds. The resulting dataset comprised 847,477 compounds for 1,254 targets from 29,091 source documents.

2.6 Addition of further target and compound information

2.6.1 Compound physicochemical descriptors

For all compounds in both the patent and non-patent datasets, physicochemical descriptors were extracted from ChEMBL via another query. Using the compounds’ “MOLREGNO”, the corresponding values from the “COMPOUND_PROPERTIES” table for molecular weight (“MW_FREEBASE”), polar surface area (‘PSA”), number of rotatable bonds (“RTB”), number of hydrogen bond donors and acceptors (“HBD” and “HBA”) and AlogP (“ALOGP”) were retrieved. For a fraction of compounds, this information was not or only partly available in ChEMBL.

2.6.2 Target information from IDG/Pharos and activity annotation

Information on the target development level and IDG family of each target was downloaded from the Pharos web interface[11] (https://pharos.nih.gov/), and added to the datasets.

For those compound-target entries where the bioactivity was available as a pChEMBL value (provided by the ChEMBL team to standardize bioactivity across assays: defined as -log(molar IC50, XC50, EC50, AC50, K_i, K_d or potency)[28]) , a designation as “active” or “inactive” was added based on the same criteria used by the IDG initiative. These differ by protein family and are based on a cutoff of 90% of approved drugs in that family showing a higher activity[5]. They are listed in Table 5.

Table 5: Activity cutoff values from IDG.

Protein family	Activity cutoff	Corresponding pChEMBL value
Kinases	≤ 30 nM	≥ 7.52
GPCRs	≤ 100 nM	≥ 7
Nuclear receptors	≤ 100 nM	≥ 7
Ion channels	≤ 10 µM	≥ 5
Other	≤ 1 µM	≥ 6

2.7 Comparison of patent and non-patent data and division into subsets

A comparison of the targets for which bioactivity data was found in patents and non-patent sources (in accordance with the criteria outlined above; 1,251 targets in total) reveals that for 119 targets, there is only data derived from patents. While there may be some data from non-patent sources with another, less strictly defined target type, or an assay type outside the scope of this project, for these targets, patents currently could be considered the most important source of bioactivity information in ChEMBL. However, only 76 of them had bioactivities given in the form of pChEMBL values, which allow an easy (albeit rough) classification into active and inactive compounds via cutoff values, as was done throughout this project by using the protein family-based activity thresholds used by the IDG initiative, shown in Table 5. 54 targets in this dataset were found to have (by definition patent-exclusive) active compounds.

Of the 1,251 targets with bioactivity data from both patents and non-patent literature, two targets had only patent compounds where no chemical structures (SMILES) were available in ChEMBL; all compounds lacking a SMILES were dropped from both the patent and non-patent datasets at this point.

For evaluating the contribution of patents in the form of patent-exclusive compounds, the compounds annotated for each target from the two data sources were compared. 1,229 of those targets where non-patent data was available had at least one patent-exclusive compound. The same comparison was made for the active compounds for each target, where 725 targets were found to have actives found only in patents. A schematic of the datasets mentioned here and later in the results section is shown in Figure 3.

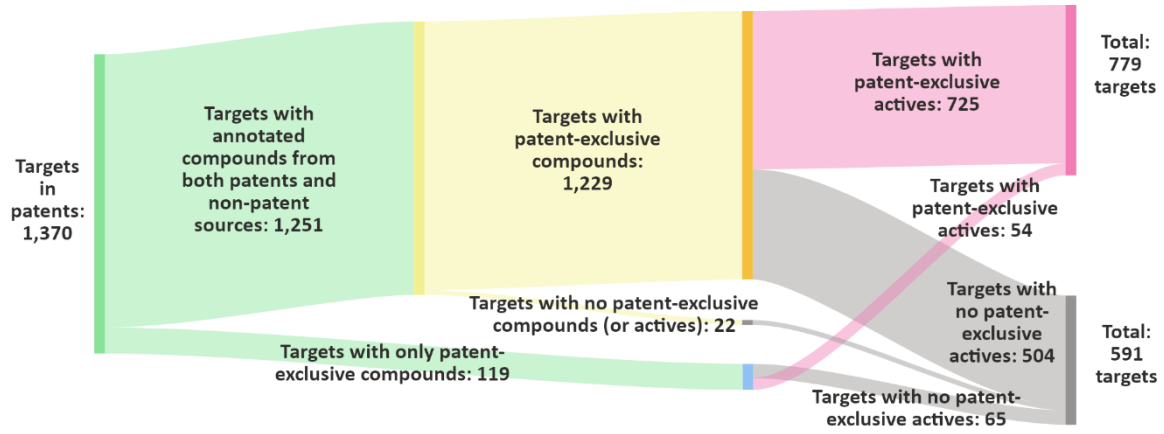


Figure 3: Schematic diagram of the subdivision of the dataset during processing. The subsets are used throughout the analysis to compare between patent and non-patent data.

2.8 Similarity/diversity calculations

Towards the goal of evaluating the value added by patent bioactivity annotations in ChEMBL, the first step was to determine the target-wise diversity of each of the subsets described in the last section. An exception is the dataset of targets that have bioactivity annotations only from patents – here, each piece of data arguably is a valuable and new addition to ChEMBL.

The calculation of the three parameters used to assess each dataset's diversity – scaffold, compound and physicochemical diversity – was adapted from González-Medina et al[32] with modifications mentioned in the individual sections below. In the article, the authors suggest a visualization of the three diversity components of a dataset in the form of a single diagram, a consensus diversity plot (CDP).

2.8.1 Scaffold diversity

There are different approaches to represent a molecule's core structure. While in the original article on CDPs, González-Medina et al are using a scaffold derived by a method described elsewhere[35] and then calculating chemotypes using Molecular Equivalent Indices (MEQI), the online version of their CDP workflow (which can be found under https://consensusdiversityplots-difacquim-unam.shinyapps.io/RscriptsCDPlots/) also references other scaffold generation possibilities such as the scaffold tree[31,36]. In this project, however, Bemis-Murcko scaffolds are used[37], which were calculated using the "MurckoScaffold" function in rdkit.Chem.

For each target in a subset, the number of different scaffolds as well as the number and fraction of compounds per scaffold were determined. An additional, easily obtainable metric of the diversity

of scaffolds for a target is the fraction of scaffolds represented by only one compound (referred to as "singletons").

The main parameter for target-wise scaffold similarity used in this project is the fraction of scaffolds representing 50% of all compounds for a target, referred to as "F_{50}". This value is calculated via linear interpolation of the cumulative fraction of compounds versus the cumulative fraction of scaffolds with numpy's "interp" function, with a higher value indicating higher scaffold diversity. Due to the nature of the calculation, the F_{50} would be 1 for targets that have only one scaffold; in these cases, it was replaced by 0 to correctly reflect the target's low scaffold diversity.

For acyclic compounds, no Bemis-Murcko scaffolds is generated due to the nature of the process. To prevent missing scaffolds from biasing the calculations of scaffold diversity, the compounds were counted as having one scaffold each in the F_{50} calculation, but they are not included in the number of scaffolds per target or singletons.

2.8.2 Fingerprint similarity

Another dimension of the diversity of compounds in a dataset, apart from their core structures as described by scaffolds, are their side chains. They are removed in the generation of scaffolds but can harbor distinctive features; to quantify these differences, fingerprints are used. There are many methods to generate fingerprints and compare them to each other[34], with different advantages and disadvantages and areas of application. This project uses MACCS keys (166 bits) and pairwise Tanimoto similarity[33], as implemented in RDKit[38]. Fingerprints for all compounds for a given target were generated using RDKit's "MACCSkeys" function, followed by the calculation of the pairwise Tanimoto similarity of all combinations of compounds of that target. The mean of all Tanimoto similarity values was taken as the parameter describing each target's compound diversity for comparison purposes.

For targets with only one associated compound, it is naturally impossible to calculate a pairwise similarity; thus, the value was set to the highest possible value of 1 in these cases, which would represent identical fingerprints if two compounds were compared.

In some cases, RDKit's "MolFromSmiles" function did not generate a mol object from a SMILES string, so no fingerprint can be generated. For the purposes of calculating the mean Tanimoto similarity across all compounds for a target, such compounds were simply excluded from this calculation, while still being included in the calculation of scaffold similarity and physicochemical similarity.

For visualization purposes, it was later decided to convert the fingerprint similarity calculated as described to fingerprint *diversity* by simply subtracting it from 1 (also referred to as Soergel distance[39]).

2.8.3 Physicochemical diversity

The procedure of calculating the physicochemical diversity of the compounds associated to a given target was adapted with very few changes from the CDP article[32]. The six physicochemical properties – molecular weight, polar surface area, number of rotatable bonds, number of hydrogen bond donors and acceptors and octanol/water partition coefficient (AlogP) – were retrieved from ChEMBL as described in section 282.6. Using scipy's "euclidian" function, the pairwise euclidian distance of all compounds of a given target was calculated, and the mean of these values was used for comparisons.

Similar to the treatment of absent values in the calculation of fingerprint similarity described above, compounds lacking any one of the physicochemical descriptors were excluded from the calculation, but not from that of the other two parameters (scaffold and fingerprint similarity). For targets that were left without any compounds after that step, the diversity value was set to "NaN"; targets with only one compound and none to compare it to were given a diversity value of zero.

2.9 Calculation of novelty values

For the evaluation of novelty of patent-exclusive compounds, similar descriptors to those used for assessment of diversity were implemented, with the important distinction that pairwise comparisons for each target were made for compounds between two datasets (patent-exclusive and non-exclusive), rather than within the same dataset. Non-exclusive compounds may either be published solely outside of the patents (so far) annotated in ChEMBL or may occur both in patents and non-patent literature.

2.9.1 Scaffold novelty

The F_{50} calculation used for intra-dataset scaffold diversity could not be adapted to inter-dataset comparisons. Instead, in addition to the values for singleton scaffolds and the number and fraction of compounds for which no scaffold could be generated which were similarly found for diversity calculations, the fraction of patent-exclusive scaffolds was used as the main descriptor of scaffold novelty.

2.9.2 Compound novelty

A target's compound novelty was calculated as the mean pairwise Tanimoto distance between the MACCS fingerprints of each patent-exclusive compound and non-exclusive compound for that

target. To emphasize novelty instead of similarity, this value was subtracted from 1, similarly to what was done for diversity. A high value was thus indicative of high novelty.

2.9.3 Physicochemical novelty

For physicochemical novelty as well, the physicochemical descriptors of each of a target's patent-exclusive compounds were compared with those of each non-exclusive compound, and the mean of all target-wise Euclidian distances was taken as the target's novelty value. Compounds missing one or more descriptor values were handled as for the diversity calculations, by excluding them.

3 Results and discussion

3.1 Understudied targets in ChEMBL

ChEMBL is one of the sources the IDG initiative uses to categorize targets, providing data on ligand bioactivity, and the information it holds and continues to collect on understudied targets directly impacts target development levels. A brief investigation into the current distribution of information across target families and development levels was performed to assess current information gaps and potentially identify strategies to alleviate them.

The 20,412 proteins listed by Pharos are divided into 10 families and one "other" category. These are not all equally well studied, as shown in Figure 4 and Table 6. For example, while there are no nuclear receptors among Tdark targets, all of the 168 nuclear receptor family members are either Tbio (14.8%), Tchem (42.2%) or even Tclin (42.8%). On the other hand, only one out of 438 olfactory GPCRs (oGPCRs) has made it to the Tchem stage of development. Apart from oGPCRs, the percentage of understudied targets from the Tdark and Tbio categories is highest among transcription factors (96.4%) and "other" proteins (94.4%), which are also the most numerous categories. After nuclear receptors, kinases, ion channels and GPCRs have the lowest fraction of Tdark/Tbio targets.

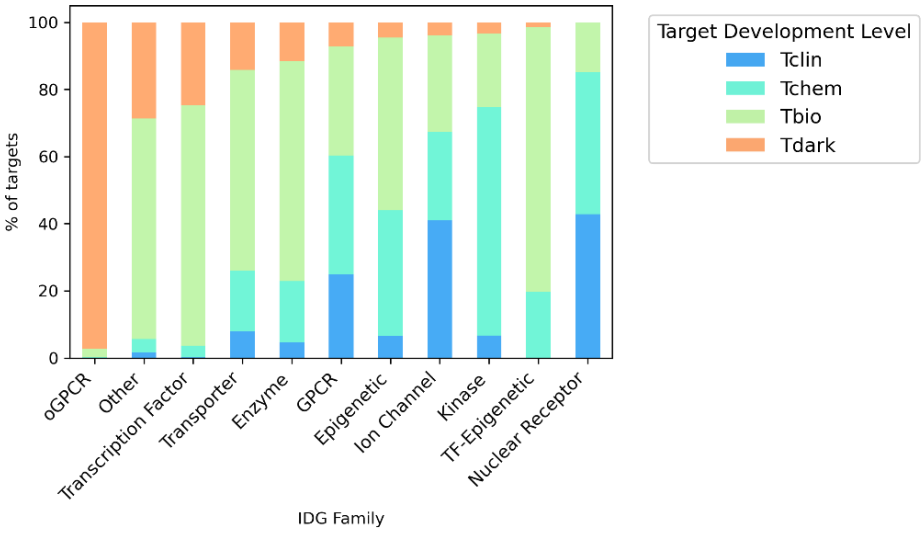


Figure 4: Target Development Level distribution among the different target families.

Table 6: Numbers of targets per family and TDL.

		Target Development Level			
IDG Family	**Total targets**	**Tclin**	**Tchem**	**Tbio**	**Tdark**
oGPCR	438	0	1	11	426
Other	18,996	313	750	12,493	5,440
Transcription Factor	2,744	6	94	1,965	679
Transporter	775	62	140	463	110
Enzyme	7,765	362	1,418	5,085	900
GPCR	747	186	265	243	53
Epigenetic	504	33	189	259	23
Ion Channel	881	361	232	254	34
Kinase	1,773	119	1,206	390	58
TF-Epigenetic	71	0	14	56	1
Nuclear Receptor	168	72	71	25	0

These differences between families are likely due to a combination of various factors. Apart from those resulting from unequal scientific interest, members of some families may be less likely to have a reported function (a prerequisite for Tbio), offer fewer possibilities for ligand binding (hampering progression to Tchem) or their modulation may not impact disease or be fraught with side effects (preventing the target from becoming Tclin with the approval of a drug).

To evaluate the level of information ChEMBL 31 contains on understudied targets, bioactivity data was retrieved for all targets designated as Tdark or Tbio in Pharos.

3.1.1 Data on Tdark and Tbio targets in ChEMBL 31

The query for bioactivity information for the 5,679 Tdark targets listed in Pharos returned a dataset of 1,988 unique compounds on only 72 targets (Table 7). Considering that these are targets about which very little knowledge exists, this is not surprising. Indeed, those targets that *are* present in ChEMBL may already be on the verge of being promoted into a higher Target Development Level. However, it also means that for more than 98% of Tdark targets, there is no bioactivity data at all available in ChEMBL.

Apart from epigenetic transcription factors (TF-Epigenetic), whose sole Tdark target is covered in ChEMBL, the highest coverage of Tdark targets is achieved for GPCRs, ion channels and kinases, in that order. These three key druggable protein families are the focus of the understudied target priority list by IDG[40] (the current and past versions of which can be found in https://github.com/druggablegenome/IDGTargets) and were the only families surveyed by IDG until their scope was expanded to include the entire human proteome, likely explaining their comparatively high coverage.

Table 7: Coverage of Tdark targets in ChEMBL31. There are no Tdark nuclear receptors currently listed in Pharos.

IDG Family	Pharos (Tdark)	ChEMBL	% in ChEMBL
TF-Epigenetic	1	1	100
Epigenetic	13	0	0
Ion Channel	19	4	21
Kinase	19	7	37
GPCR	30	17	57
Transporter	79	1	1
oGPCR	409	1	0.2
Transcription Factor	414	2	0.5
Enzyme	510	13	3
Other	4,185	26	0.6

For the comparatively better-studied Tbio targets, bioactivity data for 358,094 compounds on 1,742 targets was retrieved, representing around 14% of all 12,058 Tbio targets. By definition, these targets lack an active ligand, while there is some amount of functional information on them and there may also be antibodies available. It must be emphasized that this preliminary investigation provides no information on the depth of the bioactivity data for each target – one compound is enough to mark a target as "covered" here, so these findings must be taken with caution and are only to be used as an overview.

The number and fraction of covered Tbio targets by family are shown in Table 8. While all seven nuclear receptors have bioactivities in ChEMBL, no information is available for the 11 olfactory GPCRs. Again, GPCRs are best represented in ChEMBL with more than 71%, followed by kinases with 55%. Tbio ion channels (32%) are slightly surpassed by epigenetic targets (37%).

Table 8: Coverage of Tbio targets in ChEMBL32.

IDG Family	Pharos (Tbio)	ChEMBL	% in ChEMBL
Nuclear Receptor	7	7	100
oGPCR	11	0	0
TF-Epigenetic	29	7	24
GPCR	103	74	72
Ion Channel	108	34	32
Epigenetic	133	49	37
Kinase	161	89	55
Transporter	278	47	17
Transcription Factor	940	73	8
Enzyme	2,725	569	21
Other	7,563	793	11

Of the 111 Tbio and Tdark proteins on the IDG priority list (2022 version), bioactivities for 67 were retrieved from ChEMBL. As can be seen from Table 9, GPCRs are relatively well covered, while a large portion of the listed kinases and ion channels have no bioactivities in ChEMBL. The Tdark targets on the IDG priority list have similar coverages as for Tbio, suggesting that the ongoing annotation efforts for priority targets have indeed been stimulated by their inclusion on the list.

Table 9: Coverage of targets from the understudied target priority list.

	Tbio			Tdark		
IDG Family	**IDG22 list**	**ChEMBL**	**% in ChEMBL**	**IDG22 list**	**ChEMBL**	**% in ChEMBL**
GPCR	48	41	85	20	17	85
Kinase	33	19	58	15	6	40
Ion Channel	30	7	23	13	3	23

It is important to keep in mind that even those proteins that have annotated bioactivities are lacking any ligands with activities exceeding the IDG cutoff (at least as of the last recalculation of the TDL), as that would move them to the Tchem category. Thus, more data on these targets is still needed.

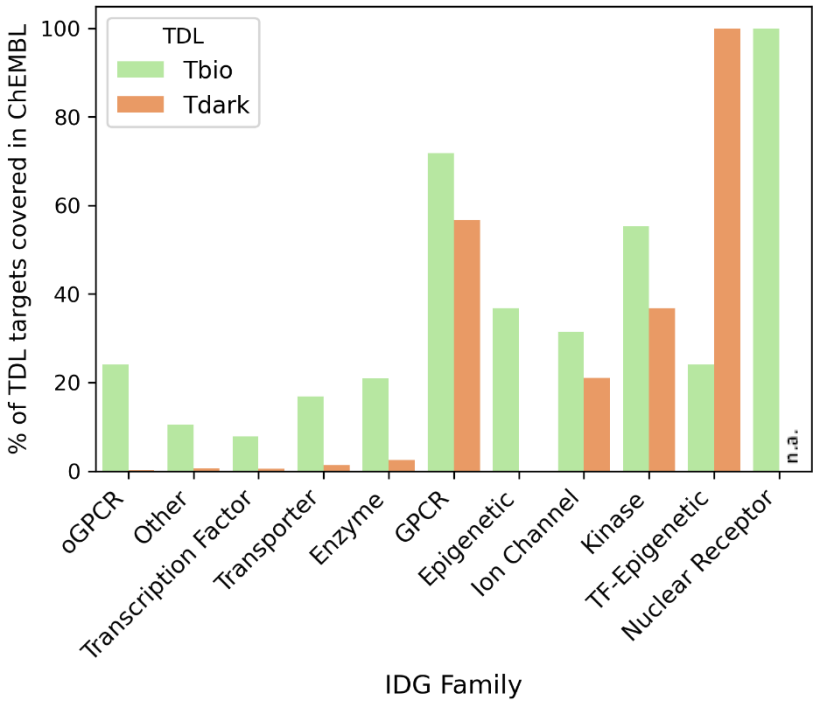


Figure 5: Coverage of understudied targets from different families in ChEMBL31. Note that there are no Tdark nuclear receptors currently listed in Pharos, thus the absence of Tdark nuclear receptors in ChEMBL does not in fact represent missing targets.

Figure 5 summarizes the above findings and depicts the percentage of Tbio and Tdark targets of each family for which there is data available in ChEMBL. Only for nuclear receptors, there are no understudied targets without data and coverage could be said to be complete; however, it must be noted that no member of that family has been assigned to the Tdark category. The complete coverage of Tdark TF-Epigenetic targets is, likewise, due to the only such target being present in ChEMBL. Both for Tdark and Tbio, the families most in need of annotations (if regarded from the point of percent of understudied targets covered in ChEMBL) are transcription factors, transporters, enzymes and oGPCRs as well as the "other" category, which is the most numerous one for both TDLs. In addition, none of the 13 Tdark epigenetic targets has bioactivity annotations in ChEMBL.

3.2 Patent bioactivities for human targets in ChEMBL

The query and processing outlined in the methods sections 2.3 and 2.4 revealed that ChEMBL 31 contains patent bioactivity data from binding or functional assays for 1,370 human targets (restricted to single proteins or protein complexes), derived either from ChEMBL's own extraction efforts, or from data exchange with BindingDB[29]. Interestingly, two of these proteins have not been assigned a Target Development Level yet: UniProt accessions Q9BRH5 (Diacylglycerol O-acyltransferase 1, DGAT1) and Q6L5J4 (Formyl peptide receptor-like 2, FML2). Both are currently unreviewed (only computer-annotated) in UniProt and therefore not included in Pharos, which contains only manually reviewed proteins[11], and each is annotated from a single patent (US-8703761-B2 and US-20120115841-A1, respectively, granted in 2014 and 2012).

Of the other 1,368 targets with patent bioactivity annotations in ChEMBL 31, 320 are in the Tclin category, and the majority (885 targets) in Tchem. Considering that bioactivities included in patents are likely to include active compounds, and the TDL designation is partly based on ChEMBL data, it is not surprising that there is a bias towards these TDLs. However, 157 Tbio and 6 Tdark targets are also found. There are several reasons why their annotations may not have led to an upgrade in TDL. Their bioactivities are not expressed as pChEMBL values (as in the case where it is only mentioned in the patent text that they do or do not exhibit activity), they do not meet the cutoff for actives, or the targets would be upgraded to Tchem at the next recalculation of TDLs.

It may also be the case that the activities in ChEMBL are for a complex containing the protein rather than the single protein version of the target. An example is UniProt accession P08648, Integrin alpha-5. As its IDG family designation is "other", meaning it does not fit into any of the other 10 families, the activity cutoff of ≤ 1 μM corresponds to a pChEMBL value of 6. Nevertheless, 16 compounds with activities greater than 7 were annotated from the patent US-8501787-B2 – but only for the Integrin alpha-5/beta-1 complex. It is important to keep the distinctive target types and how they are handled in ChEMBL and Pharos in mind.

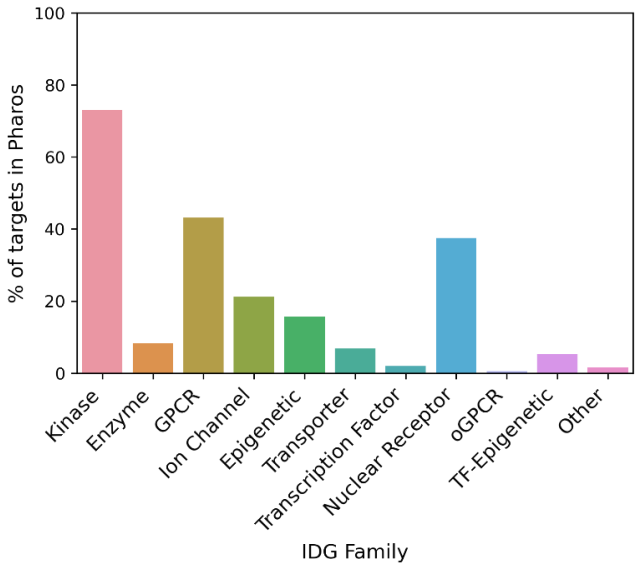


Figure 6: Patent bioactivities present in ChEMBL 31 for each target family.

Figure 6 shows the percentage of targets from each family for which patent bioactivities are present in ChEMBL 31. Most kinases (73%) and a large portion of GPCRs (43%) and nuclear receptors (38%) have received such annotations, but other families are not as well represented. In addition, the overwhelming majority of targets from any family with patent annotations belongs to either Tclin or Tchem, while Tbio and Tdark targets are distinctly underrepresented (Figure 7). Table 10 summarizes the number of targets per family and TDL for which patent bioactivities have been added to ChEMBL.

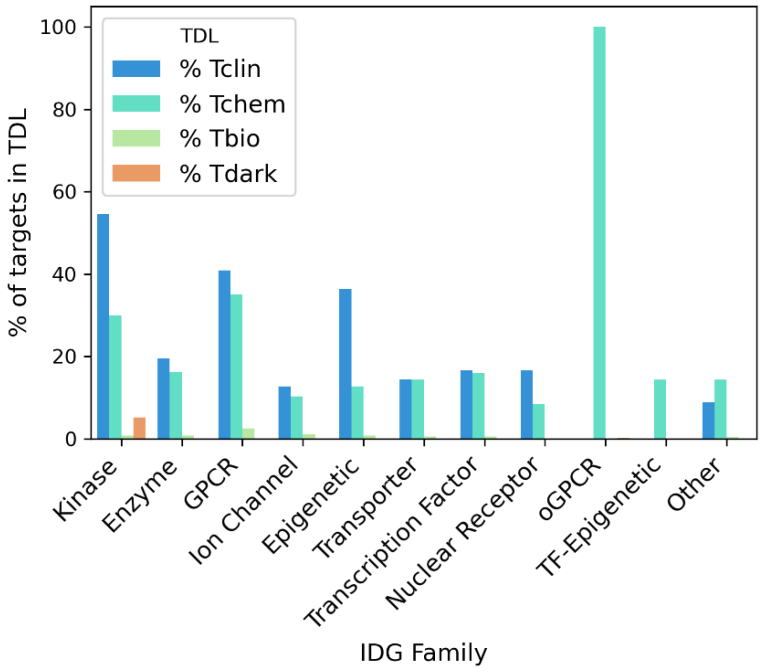


Figure 7: Patent bioactivity data in ChEMBL31 for each family, split by TDL. Note that there is only one Tchem oGPCR, which is covered in ChEMBL, hence the 100% on that bar.

Among the group of targets which were grouped together as "other" (the by far largest one, counting 12,265 members), only the low fraction of 1.6% has annotations from patents in ChEMBL. After oGPCRs, this group also has the largest fraction of understudied targets in general (see Figure 4), and the fraction of Tdark and Tbio targets from this group that is present in ChEMBL, regardless of data source, is at the low end (compare Table 7 and Table 8), so it is unsurprising that this is also seen when examining patent data only.

Table 10: Number of targets for which patent data has been incorporated into ChEMBL 31, split by TDL.

IDG Family	Total targets in Pharos	Tclin	Tchem	Tbio	Tdark
Kinase	635	65	362	3	3
Enzyme	4,141	71	230	43	1
GPCR	406	76	93	6	0
Ion Channel	344	46	24	3	0
Epigenetic	242	12	24	2	0
Transporter	472	9	20	3	0
Transcription Factor	1,400	1	15	12	0
Nuclear Receptor	48	12	6	0	0
oGPCR	421	0	1	0	1
TF-Epigenetic	38	0	2	0	0
Other	12,265	28	108	54	1

Of the 318 targets on the IDG priority list, 131 have annotations from patents. Again, kinases are best covered, with 80 of the 87 Tchem, 12 of the 34 Tbio and 3 of the 15 Tdark kinases on that list having patent bioactivity annotations (Table 11). In contrast, no GPCRs or ion channels currently designated as Tbio or Tdark have been annotated from patents, although some may have been elevated to Tchem after (or even due to) annotation.

Table 11: Patent annotations for targets from the IDG priority list. Numbers in parentheses are the number of targets from that family and TDL on the list.

IDG Family	Tclin	Tchem	Tbio	Tdark
GPCR	0 (0)	34 (48)	0 (50)	0 (20)
Ion Channel	1 (16)	1 (4)	0 (31)	0 (13)
Kinase	0 (0)	80 (87)	12 (34)	3 (15)

3.2.1 Diversity of bioactivity annotations from patents

The purpose of this section is to attempt to quantify the knowledge gained from the annotation of compound bioactivities from patents (either by the ChEMBL team itself, or from data exchange with BindingDB) by calculating the diversity of the datasets for each target. It is important to keep in mind that this does not necessarily reflect the overall level of information about each target in ChEMBL, as patents are only one of the sources of data, and one of far smaller coverage than scientific literature at this point in time.

In this first analysis, no distinction of active and inactive compounds was made, as its aim is to describe the overall diversity of compounds with bioactivity data for each target. This is following the notion that information on inactive compounds is just as valuable as on active ones, and that even those where activity is not readily available (i.e., where activity was not available as pChEMBL value) may be of interest. The diversity of actives and inactives for each target is discussed in later sections.

Of the 1,370 targets with patent annotations in ChEMBL, two were excluded from further analysis as none of the compounds for them had a SMILES deposited and thus the diversity calculations could not have been performed. The number of compounds annotated from patents for each target varies widely, as shown in Figure 8. While 89 targets are annotated only with a single compound originating from patents in CHEMBL 31, and most targets listed are annotated with between two and fifty compounds, 49 targets had over 1000 unique compounds annotated. The highest number of annotated compounds (4,314) was found for the Tyrosine-protein kinase JAK2 (Tclin, O60674), with annotations originating from 45 different patents.

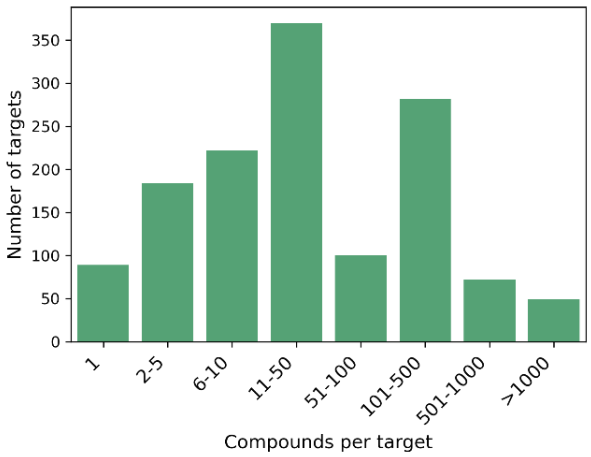


Figure 8: Number of compounds annotated from patents per target.

The main measure of scaffold diversity for each target used in this thesis is the fraction of scaffolds that encompass 50% of its compounds, the F_{50}. A high F_{50} value indicates high scaffold diversity, as a large number of scaffolds is required to cover half of the compounds. The maximum value the F_{50} can assume is 0.5, where each compound has a different scaffold. In the special case where there is only one scaffold for a target, the calculation leads to a value of 1, which does not reflect the real lack of diversity, and is consequently replaced by 0. Another interesting angle is to examine the fraction of "singletons": Scaffolds that are only present in one compound per target. If singletons are making up a large portion of a target's scaffolds, it suggests there are unexplored possibilities within these scaffolds.

The diversity among the compounds for a given target was calculated as mean pairwise Tanimoto similarity between their MACCS keys (166-bit), converted to Soergel distance by subtracting it from 1. A high value thus indicates high fingerprint diversity. Conversely, targets with only one associated compound, or identical fingerprints for all compounds, were assigned a diversity of 0.

The third diversity measure used in this project is the mean pairwise Euclidian distance of six physicochemical descriptors (molecular weight, polar surface area, AlogP, number of rotatable bonds, number of hydrogen bond donors and acceptors) between compounds for a given target. These descriptors were retrieved from ChEMBL; however, they were not deposited for all compounds. Compounds where any of the descriptors is missing are not included in the calculation. Consequently, for some targets, fewer than two compounds with all descriptors were available and this value could not be calculated (set to NaN). For targets with only one compound, the distance was set as zero to reflect the lack of diversity. In contrast to the other two diversity descriptors, the value of the physicochemical diversity is not by its nature normalized and therefore any comparisons between targets can only be made within a dataset.

The three diversity measures are not entirely independent of each other but describe different aspects of a target's chemical space and its deficiencies. They are summarized in consensus diversity plots, where scaffold and compound diversity increase with a higher value on both x and y axes. Targets in the upper right of the plot have the highest compound and scaffold diversity among this dataset, and those in the lower left section have the lowest. The latter are consequently those targets with the greatest knowledge deficit and highest need of further annotations. Physicochemical diversity is depicted via the marker color: as this value is not bound by limits as the others, each target was assigned to one of four categories according to their quartile among all targets within a given dataset.

3.2.1.1 Overall diversity of patent bioactivity annotations

As 1,368 datapoints are too many to show individually on a single plot, a contour plot was chosen to show the areas of highest density among all targets in this dataset as an overview (Figure 9). There are three areas with an especially high number of targets: Those with a compound and scaffold diversity of 0 due to only having one annotated compound (lower left corner); those with the maximum scaffold diversity of 0.5 (upper edge), and those with rather low scaffold and medium fingerprint diversity (near the bottom, middle of the plot area). Physicochemical diversity could not be incorporated in this representation, but is included in those plots depicting single families, further below.

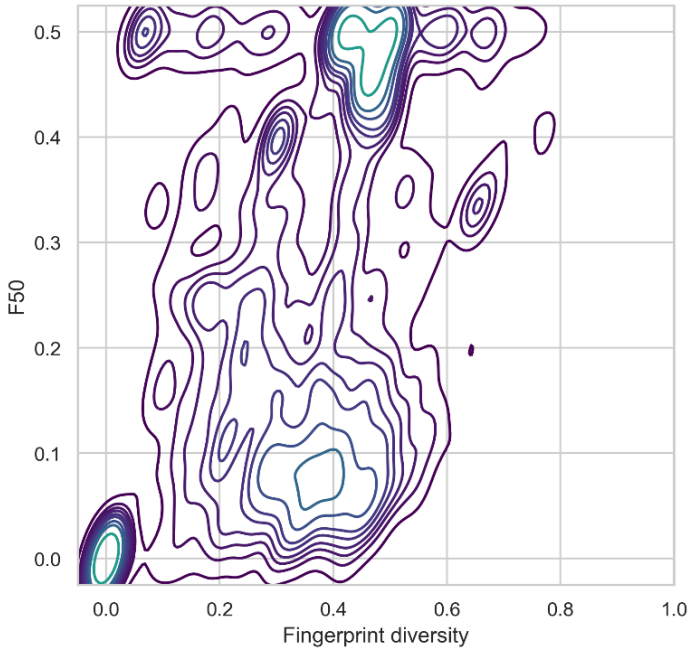


Figure 9: Contour consensus diversity plot for all targets with patent annotations, showing their distribution along the compound (x-axis) and scaffold (y-axis) diversity parameters. Physicochemical diversity is not shown in this representation; the color indicates only the density of datapoints in the plot, with the green hue indicating higher density.

Among the 1,368 targets in this dataset, 318 had a singleton fraction of 1. While 85 of them had only one compound annotated, a high singleton fraction does not necessarily equal low diversity: One target, DNA replication ATP-dependent helicase/nuclease DNA2 (Tbio; UniProt: P51530), lists 38 compounds, all with a distinct scaffold. 75 targets had at least one compound where no scaffold was generated.

For 114 of the 1,368 targets, the F_{50} value was 0 due to only one scaffold being found for all compounds. As mentioned before, this often is accompanied by a low number of compounds in total (in this dataset, up to 24) and indicates very low scaffold diversity. Conversely, 266 targets had the highest possible F_{50} of 0.5. These had either a singleton fraction of 1, or they had two scaffolds with equal distribution of compounds, or one of their two total scaffolds was a singleton.

Apart from these special cases, the target with the highest scaffold diversity was Vasoactive intestinal polypeptide receptor 1 (Tchem; P32241) with an F_{50} of 0.484. All of its 33 compounds were annotated from a single patent (US-20160045572-A1). It may sound unusual that one patent is covering so many scaffolds, but this is the case because the compounds are peptides, for which any difference in amino acid sequence leads to a distinct scaffold. However, putting this deceptively high F_{50} value in context with the fingerprint diversity (0.043 on a scale from 0 to 1) for the same target, it is obvious that the individual compounds are quite similar to each other. The second-highest F_{50} value is reached by Mitogen-activated protein kinase 13 (Tchem, O15264), whose 22 compounds are small molecules and stem from 7 different patents.

The lowest F_{50} (0.011) in this dataset was for the Sodium/glucose cotransporter 1 (Tclin, P13866), which has a rather large number of 640 unique compounds with 93 scaffolds. The low value is due to 360 of these compounds sharing the same scaffold.

Focusing on compound diversity, the highest-ranked target in this dataset, with a fingerprint diversity of 0.825, was M-phase inducer phosphatase 2 (Tchem, UniProt: P30305). However, this value was calculated based on only one pair of compounds that originate from two separate patents and which are quite dissimilar. Excluding targets with fewer than 20 compounds, the highest diversity was that of Taste receptor type 2 member 38 (Tchem, P59533), with a value of 0.769, whose 22 compounds stem from 4 patents. Four targets, each with fewer than 6 closely related compounds, had identical fingerprints for all of them, and thus were given a diversity of 0.

118 targets either had only one compound, or fewer than two of the compounds had a full set of physicochemical descriptors in ChEMBL. This left 1,250 targets for which the physicochemical diversity values could be compared. A wide variation of values was found, ranging from 1.2 (Thrombopoietin receptor; Tclin, UniProt: P40238) to 613.5 (Dual specificity protein phosphatase 3; Tchem, P51452). Again, targets at either extreme end of the scale tend to have relatively few annotated compounds; in the case of the two abovementioned targets, it is 5 and 3, respectively.

3.2.1.2 Diversity analyses by target family

The consensus diversity plots in the following figures depict all three diversity values for targets from each family, with marker color signifying physicochemical diversity, and marker size representing the number of compounds per target (i. e., depth of data).

Figure 10 shows the consensus diversity plots for targets from the three families that were the initial focus of the IDG project: Kinases, GPCRs and ion channels. One major difference between the three families is the number of targets with patent annotations: 460 kinases, 168 GPCRs and 53 ion channels. While for GPCRs, there were some targets with rather high compound and scaffold diversities (in the upper right corner of the plot), the fingerprint diversity of kinases and ion channels was comparatively low, with maximum values around or below 0.6 – though this may also reflect an intrinsic property of members of those families. Most ion channels also had rather low scaffold and physicochemical diversity from patent annotations, while kinases appear to be split in a high and a low scaffold diversity group.

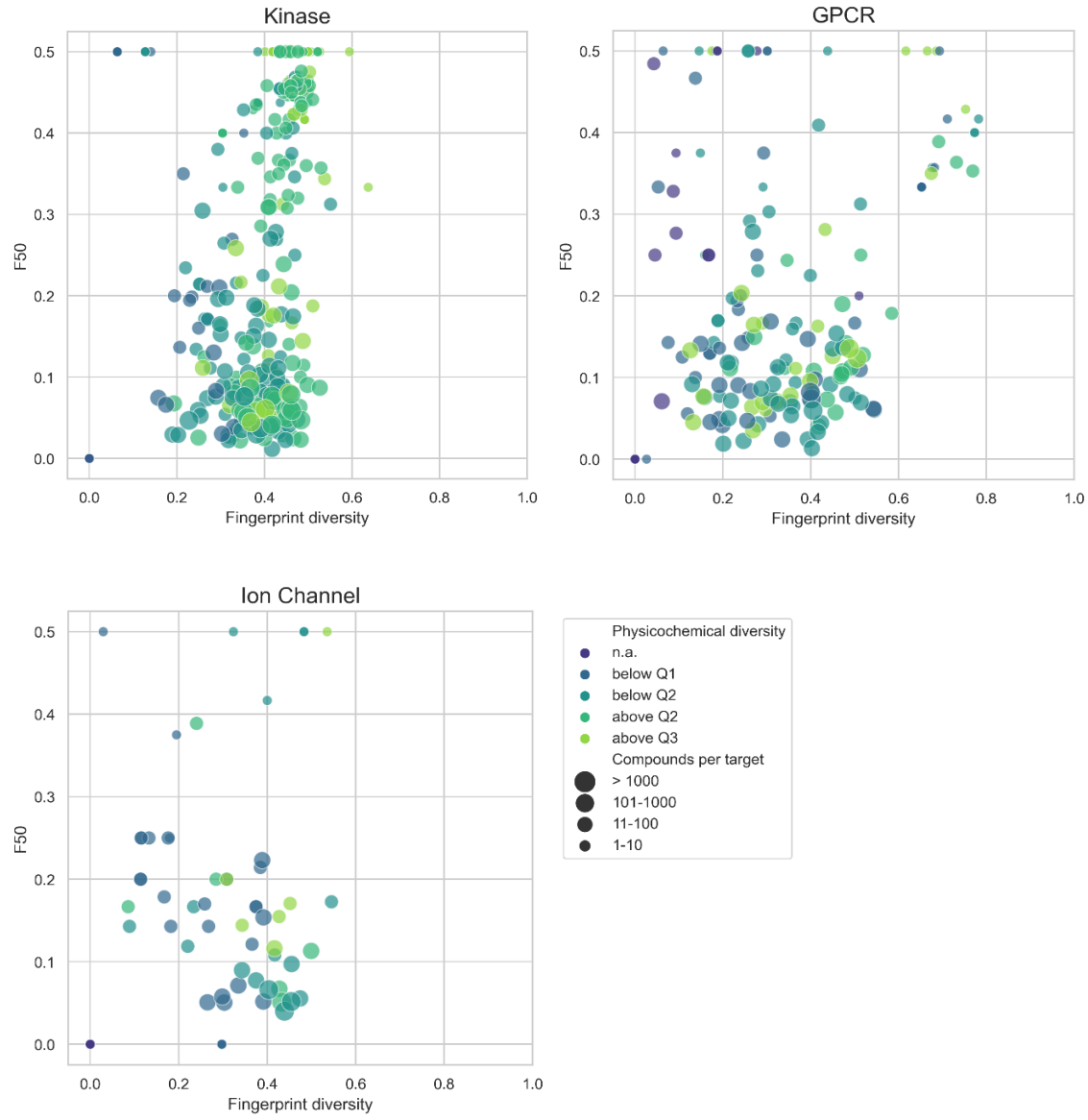


Figure 10: Consensus diversity plots of IDG priority families: Kinases, GPCRs and ion channels. The color of each marker represents a target's physicochemical diversity, while the size indicates the number of compounds annotated for each target.

Strikingly, it appears that at only targets with comparatively few compounds (up to 100) are found in the upper half of the scaffold diversity scale, while the targets with high numbers of compounds tend to have rather low scaffold diversity. That might reflect the depth of research that has been done on a target: Compounds from congeneric series and detailed SAR investigations would lead to a high fraction of compounds assigned to very few scaffolds, and thus a low F_{50} value. In the entire 1,370 target dataset, Tclin and Tchem targets make up 36% and 60% of all targets with more than 100 compounds, respectively, while 4% of Tbio and none of the Tdark targets reach that number of compounds (Table 12). That may seem a low percentage, it must be seen in the context that these understudied targets are expected to have very few, if any, annotated compounds.

Table 12: Distribution of targets from different TDLs among compound number categories.

Number of compounds	> 1000 (49 targets)	101-1000 (354 targets)	11-100 (470 targets)	1-10 (495 targets)
Tclin	26	120	114	60
Tchem	22	218	318	325
Tbio	1	15	35	106
Tdark			2	4

Two rather broad categories, each with many members, are enzymes and "other" targets, with 301 and 191 members in this dataset, respectively. As evident from the CDPs shown in Figure 11, targets in both categories span the whole range of scaffold and compound diversity values and physicochemical diversity classes. Despite these targets not being prioritized due to their family, clearly considerable contributions have been made from patent annotations. Again, targets with a higher number of compounds tend towards lower scaffold diversity; compound diversity does not appear to be as affected, with some targets with over 100 compound annotations also towards the higher end of the respective spectrum.

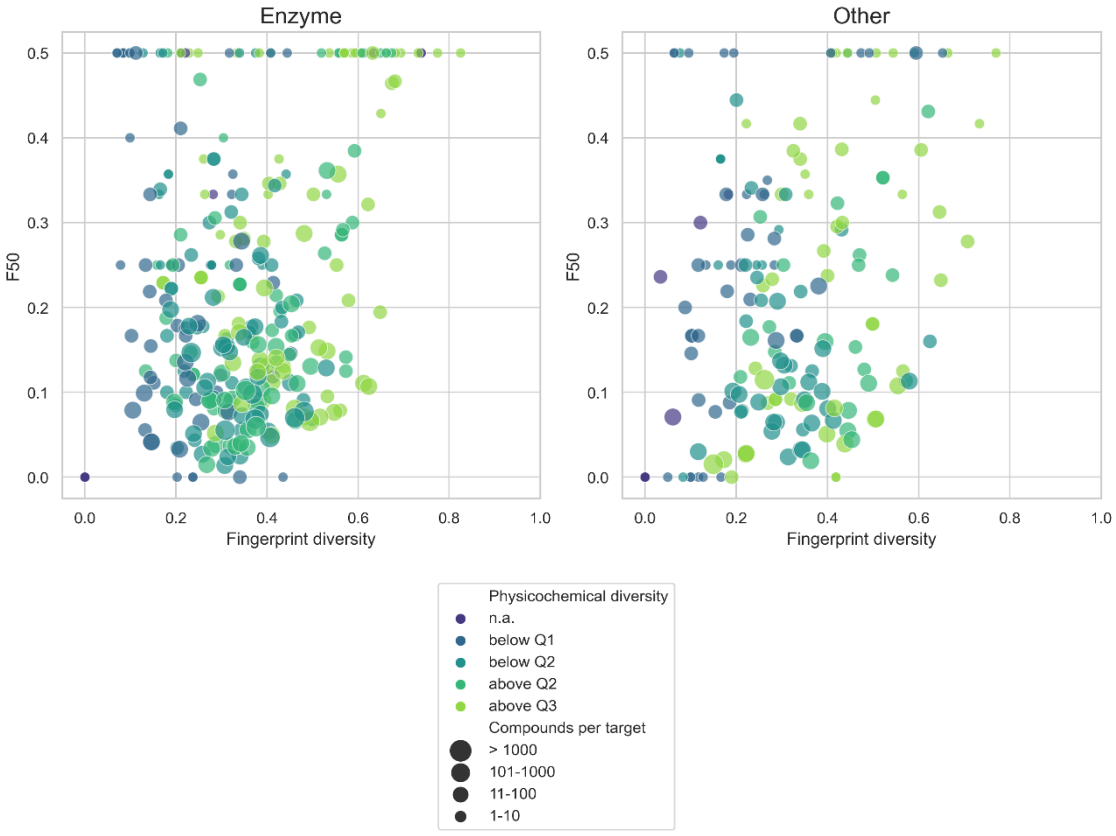


Figure 11: Consensus diversity plots for enzymes and "other" targets.

Fewer than forty respective targets were annotated from patents for epigenetic proteins, transporters, transcription factors and nuclear receptors, shown in Figure 12. In each of these families, the majority of targets does not have very diverse patent annotations, especially in terms

of compound diversity. Those targets with larger numbers of compounds again tend towards lower overall diversity values. oGPCRs and TF-Epigenetic each had only two targets with patent annotations and are not shown as separate plots.

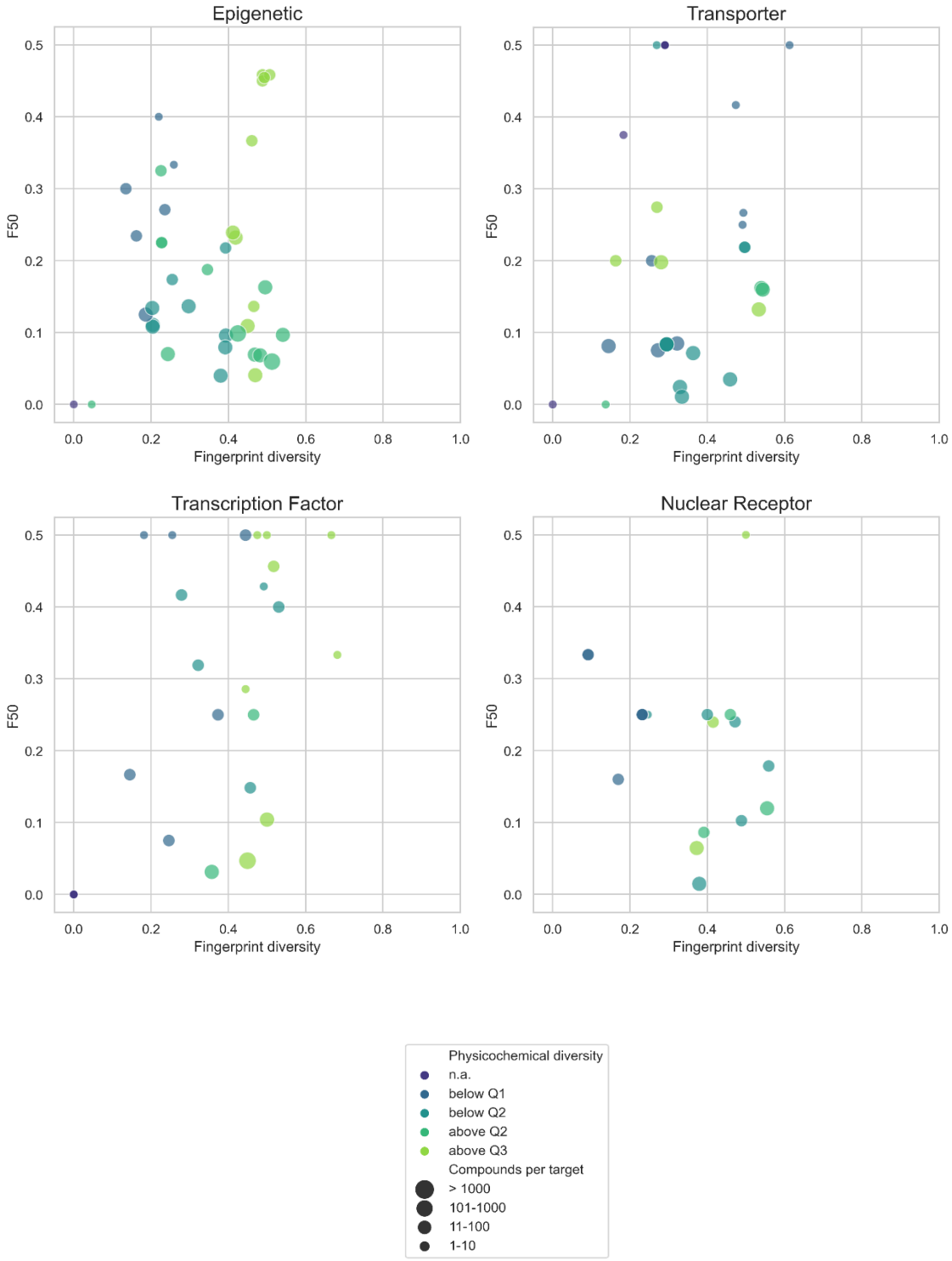


Figure 12: Consensus diversity plots for epigenetic targets, transporters, transcription factors and nuclear receptors.

To aid in prioritization of targets and to evaluate existing information, the plot can be separated into quadrants by setting thresholds at half of the maximum values on each axis. Targets above the respective threshold are designated as high diversity, those equal or under the threshold as low diversity. However, while half of F_{50} would likely be 0.25 for any dataset, provided it contains targets where each compound has a different scaffold, it must be kept in mind that the threshold for fingerprint diversity will vary with the dataset analyzed. When comparing different target families, it makes sense to apply the mean for each family separately, since not all families may have the same range of values due to intrinsic properties.

The physicochemical diversity it not given equal importance as the other values for several reasons. Firstly, it is not normalized and thus is not comparable between datasets, while compound and scaffold diversity are, and secondly, it can be assumed that some targets can accommodate compounds with a larger range of physicochemical variation than others. Thirdly, as the workflow was designed here, a portion of compounds are missing from the calculation as one or more descriptors were not available from ChEMBL. This could be remedied by an additional step that calculates the missing values, leading to a greater reliability of the results.

According to the considerations outlined above, the targets were split in four categories (Figure 13): High compound and scaffold diversity (HcHs, 421 targets); high compound and low scaffold diversity (HcLs, 340 targets); low compound and high scaffold diversity (LcHs, 153 targets); and low compound and scaffold diversity (LcLs, 454 targets).

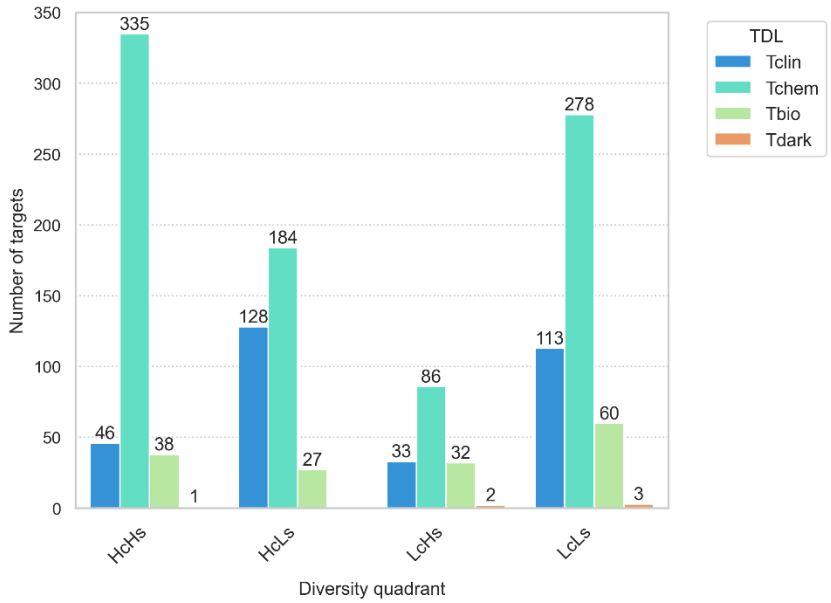


Figure 13: Distribution of targets from different TDLs across diversity quadrants. HcHs, high compound and scaffold diversity; HcLs, high compound, but low scaffold diversity; LcHs, low compound, but high scaffold diversity; LcLs, low compound and scaffold diversity.

Focusing on understudied targets from the Tbio and Tdark development levels suggests that further annotation efforts are justified: 38% of Tbio and 3 of the 6 Tdark targets are in the lowest diversity quadrant (Table 13 and Figure 14). However, clearly a lot of information has also been made available through patents, as 24% of Tbio targets and one Tdark target (Casein kinase I isoform alpha-like, UniProt: Q8N752) are found in the highest diversity quadrant. This is lower than the corresponding percentage of Tchem targets (38%), but it is important to keep in mind that since the assignment of development levels takes into account the presence of data in ChEMBL, a portion of these targets may have been elevated from Tbio or Tdark to Tchem in the course of this investigation.

Table 13: Number of targets from different TDLs in each diversity quadrant. HcHs, high compound and scaffold diversity; HcLs, high compound, but low scaffold diversity; LcHs, low compound, but high scaffold diversity; LcLs, low compound and scaffold diversity.

TDL	Total number of targets	HcHs	HcLs	LcHs	LcLs
Tclin	320	46	128	33	113
Tchem	883	335	184	86	278
Tbio	157	38	27	32	60
Tdark	6	1		2	3

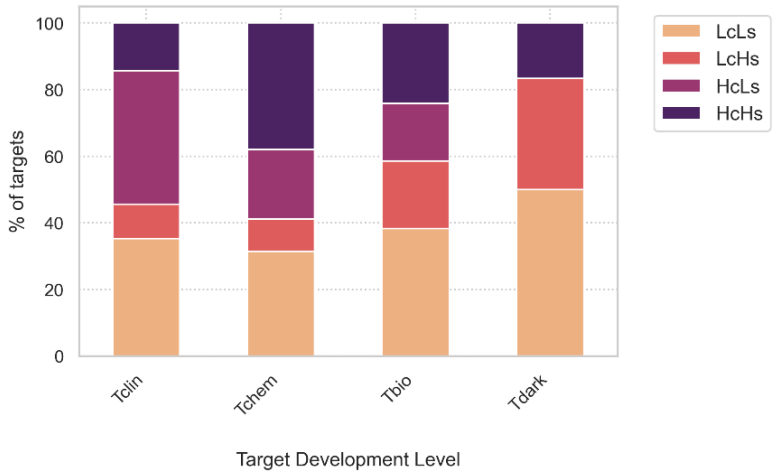


Figure 14: Fraction of each TDL's targets in the different diversity quadrants. HcHs, high compound and scaffold diversity; HcLs, high compound, but low scaffold diversity; LcHs, low compound, but high scaffold diversity; LcLs, low compound and scaffold diversity.

Figure 15 and Table 14 show how targets of different families are distributed among the four quadrants. In summary, excluding families with few members in this dataset, patent annotations yielded the highest diversity for kinases (56% of targets in the HcHs quadrant), followed by transcription factors (36%) and enzymes (20%, equal to "other" targets). However, for most families (with the exception of kinases), a third or more of their targets are in the lowest diversity quadrant. Of course, this does not take into account how much information from other sources is available in ChEMBL 31 for these targets.

Table 14: Numbers of targets in each diversity quadrant per family. HcHs, high compound and scaffold diversity; HcLs, high compound, but low scaffold diversity; LcHs, low compound, but high scaffold diversity; LcLs, low compound and scaffold diversity.

IDG Family	Total number of targets	HcHs	HcLs	LcHs	LcLs
Kinase	464	258	60	50	96
Enzyme	343	70	100	38	135
GPCR	175	27	53	23	72
Ion Channel	73	5	35	3	30
Epigenetic	38	5	16	5	12
Transporter	32	3	8	7	14
Transcription Factor	28	10	6	3	9
Nuclear receptor	18	2	9	2	5
oGPCR	2	1			1
TF-Epigenetic	2		1		1
Other	191	39	51	22	79

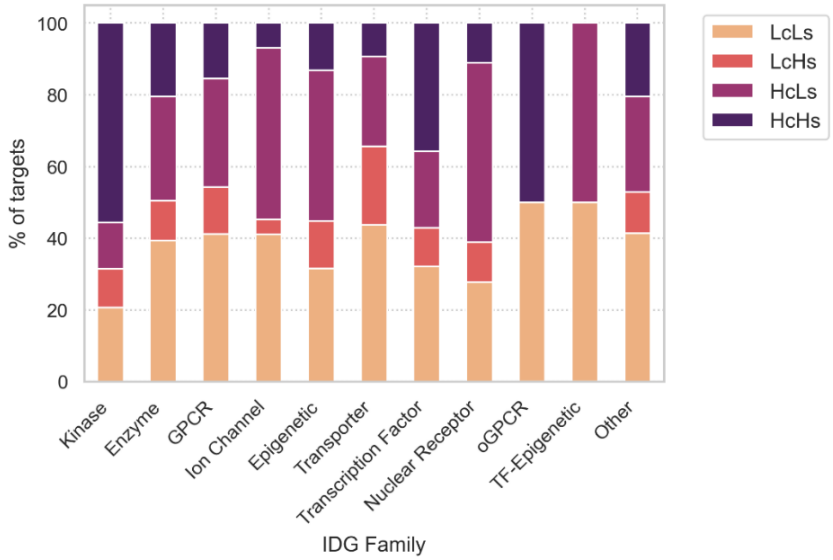


Figure 15: Distribution of targets from each family among the diversity quadrants. HcHs, high compound and scaffold diversity; HcLs, high compound, but low scaffold diversity; LcHs, low compound, but high scaffold diversity; LcLs, low compound and scaffold diversity.

Targets from the IDG priority list are a separate group of targets worth looking into in more detail, since that list is one of the resources used to prioritize patents for annotation by ChEMBL. Figure 16 shows a CDP for the 131 targets this dataset contains from that list only. Quite a few of them have relatively high diversity values, but the number of compounds tends to be rather low, and many of the higher diversity targets have only few annotated compounds.

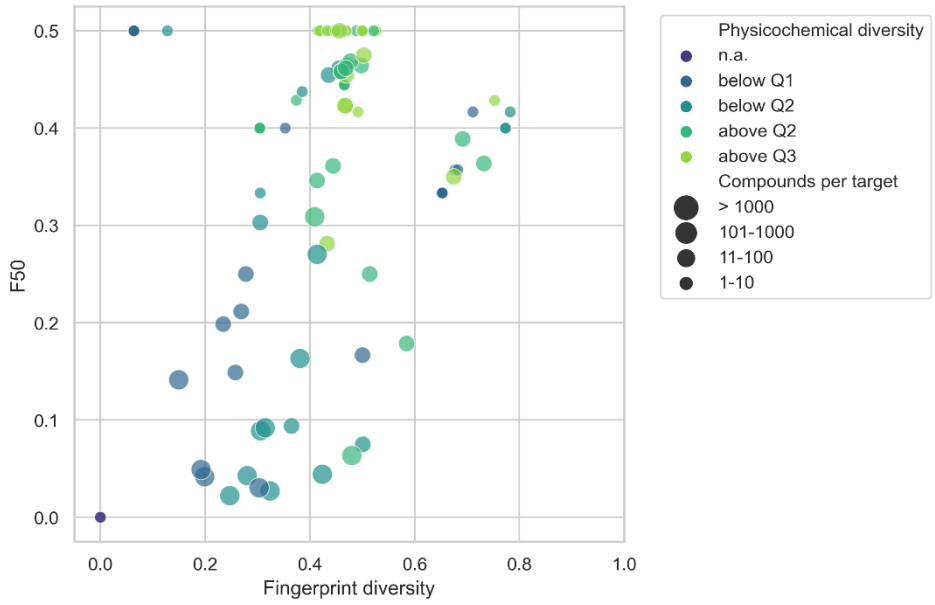


Figure 16: Consensus diversity plot for IDG priority list targets.

As detailed in Table 15, 93 (71%) of the IDG priority list targets in this dataset are found in the highest and 16 (12%) in the lowest diversity quadrant, with the other targets either showing low compound (15 targets) or scaffold diversity (7 targets). Among these prioritized targets as well, kinases have benefitted most from patent annotations, with 74 of the 95 kinases in this dataset in the HcHs quadrant. Conversely, only two ion channels out of 64 on the priority list have any patent annotations, and they each have only one annotated compound, placing them among the least diverse targets.

Table 15: Number of IDG priority targets in each diversity quadrant, split by family. HcHs, high compound and scaffold diversity; HcLs, high compound, but low scaffold diversity; LcHs, low compound, but high scaffold diversity; LcLs, low compound and scaffold diversity.

IDG Family	Number of targets	HcHs	HcLs	LcHs	LcLs
Kinase	95	74	1	14	6
GPCR	34	19	6	1	8
Ion Channel	2				2

Compared to all targets in this dataset (see Table 14), those on the IDG list appear to have a higher likelihood to be among the more diverse quadrants. However, the attention is not equally distributed among the three target families it is concerned with, with potential for improvement especially among ion channels.

As mentioned at the beginning of this chapter, two of the 1,370 targets with patent annotations were not yet registered in Pharos as they have not been manually reviewed by UniProt. Interestingly, the diversity gained from patents for these targets is quite high: one of them, Q9BRH5 (15 compounds with 14 different scaffolds), is in the HcHs quadrant, while the other (Q6L5J4; 413 compounds with 176 scaffolds) was classified as HcLs. For both, the physicochemical diversity was rather low (below Q2 compared to the other targets in the dataset).

This chapter presented a high-level view of the diversity of patent data in ChEMBL. To better illustrate the value of patent bioactivity annotation, a subset of targets for which no bioactivity data conforming to the processing criteria was annotated from non-patent literature was analyzed separately in more depth in the next chapter.

3.3 Targets with bioactivity data only from patents

For 119 of the 1,370 human targets presented above, the only source of bioactivity information were patents. As ChEMBL would otherwise not contain this data unless it were published in scientific journals – which for many compounds happens only years after their description in patents (if ever) –, the value of patent annotations is clearly illustrated by this finding.

For one of the 119 targets (Diacylglycerol O-acyltransferase 1, UniProt: Q9BRH5), its status in UniProt is "unreviewed" and it is therefore not included in Pharos yet, as discussed. Of the remaining 118 targets, the majority of 60 are in the Tchem category; 53 are Tbio, and two are Tdark (Table 16). No targets from the nuclear receptor or epigenetic transcription factor (TF-Epigenetic) families were present in this dataset. Between one and 2,012 compounds were annotated per target.

Table 16: Targets for which all bioactivities in ChEMBL stem from patent annotation.

IDG Family	Number in patent-only dataset	Tclin	Tchem	Tbio	Tdark
Kinase	2		2		
Enzyme	32		15	17	
GPCR	8		7	1	
Ion Channel	4	2	2		
Epigenetic	1		1		
Transporter	4		2	2	
Transcription Factor	11		4	7	
oGPCR	2		1		1
Other	54	1	26	26	1

Perhaps surprisingly, three targets in this set are Tclin: Sclerostin (UniProt: Q9BQB4), the potassium voltage-gated channel subfamily H member 3 (Gene symbol: KCNH3; UniProt: Q9ULD8) and the potassium voltage-gated channel subfamily A member 7 (Gene symbol: KCNA7; UniProt: Q96RP8). Usually, one would expect that a target which has reached Tclin, meaning there is an approved drug, would be well-represented with bioactivities in ChEMBL due to the extent it would have been studied. To check for missing data or a potential opportunity for improving the breadth of annotations, these targets were investigated in more detail.

For Sclerostin, a target of interest in the treatment of osteoporosis[41], all 23 ligands (16 of which meet the activity threshold of ≤ 1 µM, all of them peptides) in ChEMBL 31 indeed stem from patents (US-9708375-B2 and US-20140023653-A1). The approved drug targeting Sclerostin is also not a small molecule, but a monoclonal antibody, romosozumab (approved 2019). The 69 active ligands listed for KCNH3 in Pharos as well as ChEMBL stem from two patents (US-20080227785-A1 and US-20120088772-A1). It therefore appears that for these targets either there really are no bioactivities published in scientific journals curated by the ChEMBL team, or that these have escaped the attention of all data sources utilized by the IDG. In contrast, for KCNA7, ChEMBL

contains only one active compound from the patent US-20060079535-A1, while Pharos lists another 5 active ligands whose sources are Guide to Pharmacology[42] and PubChem[43]. These bioactivities were published in two articles from 1998 and 2002 first describing and characterizing the protein (PubMed IDs: 9488722 and 11896454, published in Journal of Biological Chemistry and European Journal of Human Genetics, respectively). This is not the main type of article that would be curated for ChEMBL, which likely explains the discrepancy of the IDG data sources.

The two Tdark targets in this dataset are the olfactory receptor 5K1 (UniProt: Q8NHB7) and the putative short transient receptor potential channel 2-like protein (Q6ZNB5). Each was annotated from one patent (WO-2013171535-A2 and US-20140235619-A1, respectively); while Q8NHB7 was the sole target its patent was focused on, for Q6ZNB5, there were three Tchem and two Tclin targets annotated from the same patent. Neither of the two targets has bioactivity annotations that provide pChEMBL values, as they are expressed as % inhibition, or only mentioned in the text without an attached value. Nevertheless, the comment field in ChEMBL sometimes provides activity information. This information, however, were not incorporated into the analyses in this project, and is not considered for inclusion of ligands in Pharos. As this represents already existing accessible information on active ligands, it may only be a matter of time until such targets can progress to Tchem.

Of the 53 Tbio targets only annotated from patents, the majority (26 targets) stems from the “other” category, while 17 are enzymes, 7 transcription factors, two transporters and one GPCR. Only 14 of these targets have bioactivities that can be expressed as pChEMBL values; none of them passes the cutoff values that would allow them to become Tchem, although some are very close. For the other Tbio targets that do not have any pChEMBL values attached, more information is found in the comment field, similar as for the Tdark targets discussed above.

Tchem targets in this dataset are, as expected, better represented regarding pChEMBL data. 58 of the 60 targets in that category have annotations with pChEMBL values, and for 54 of them, these meet the respective IDG family cutoff criteria to be counted as actives. The two targets missing pChEMBL values are the Sodium-dependent phosphate transport protein 2B (Gene symbol: SLC34A2; UniProt: O95436) and Inositol-trisphosphate 3-kinase B (Gene symbol: ITPKB; UniProt: P27987). For ITPKB, the active ligand from Pharos or the article it appears in are not listed in ChEMBL, though there are several other, structurally related compounds designated as active in the comment field. A look into ChEMBL reveals that for SLC34A2, the active ligand listed in Pharos is present in ChEMBL and was annotated from scientific literature; it was measured in an assay categorized as ADMET, which was excluded during processing for the analyses presented here. While exclusion of this assay type may be justified for most targets, this decision

could potentially lead to targets with direct ADMET relevance to appear less well represented than they are.

Seven of the 119 targets are on the IDG priority list, most of them Tchem: the taste receptor type 2 members 10, 19, 20, 40 and 42 (UniProt: Q9NYW0, P59542, P59543, P59535, Q7RTR8); as well as the serine/threonine protein kinase 19 (P49842), and the Tclin target KCNA7 already described above.

As mentioned, the targets have large variations in compound count: 28 of the 119 targets (23.5%) only have one compound annotated; 35 more have between 2 and 10 compounds. In contrast, 10 targets have more than 100 listed compounds; one of them (Tissue factor pathway inhibitor, Tchem; UniProt: P10646) counts a total of 2,010 associated compound bioactivities from 2 patents.

The diversity results for these targets are described in detail in the following sections.

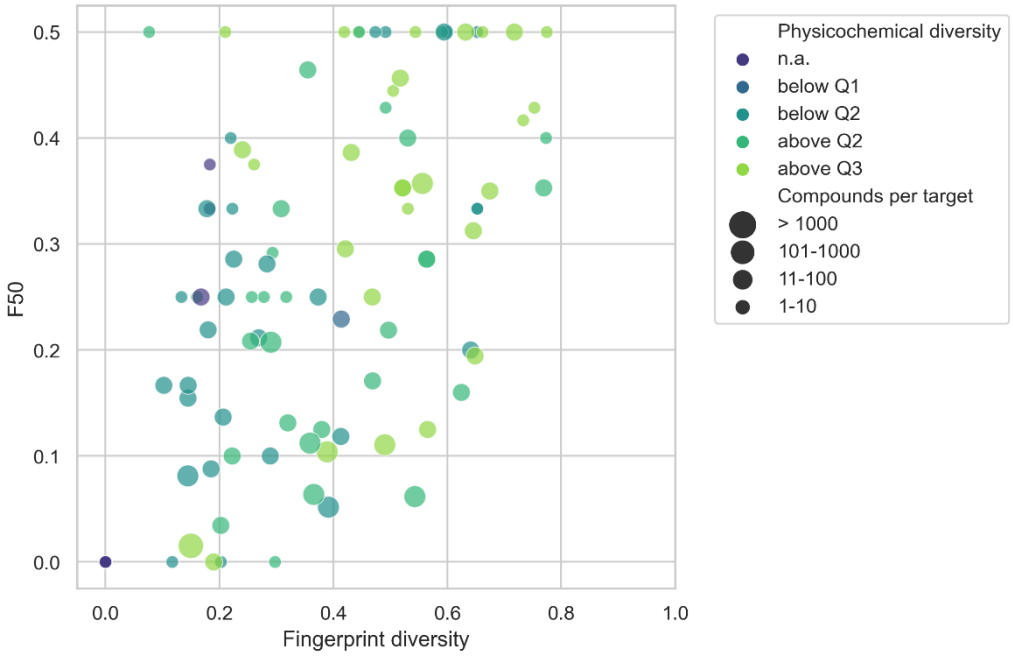


Figure 17: Consensus diversity plot for targets with bioactivities only from patent annotations.

Figure 17 shows the CDP for all 119 targets whose bioactivities stem solely from patents. Here, 39 targets have a singleton fraction of 1, indicating that each of their scaffolds only covers one compound. 25 of these targets only have one compound in total, but there are also targets with only singleton scaffolds that have a decent number of compounds, such as the DNA replication ATP-dependent helicase/nuclease DNA2 (Tbio; UniProt: P51530), which has 38 distinct singletons. 18 targets have compounds for which no scaffold could be generated, which are then counted as having a distinct scaffold when calculating the F_{50} value. 32 targets have only one scaffold and thus a calculated F_{50} of 1 which was set to 0 to reflect the absence of scaffold diversity.

Apart from those targets with a singleton fraction of 1, which also leads to a high F_{50}, targets with moderate numbers of compounds and scaffolds rank highest in scaffold diversity: Among the eleven targets with the highest F_{50} (some of which have equal values), seven have less than ten compounds, and only one has more than 20. None of them has a singleton fraction lower than 0.5.

The compound diversity is calculated from the mean pairwise Tanimoto distance of MACCS keys; thus, the 28 targets which only have one associated compound have a fingerprint diversity of 0. For the remaining 91 targets, the value varies between 0.77 and 0.08. The highest diversity was found for Mesoderm-specific transcript homolog protein (Tbio; Q5EB52), for which only 3 compounds were annotated from the patent US-20120270250-A1. However, they indeed share very little similarity, as shown in Figure 18.

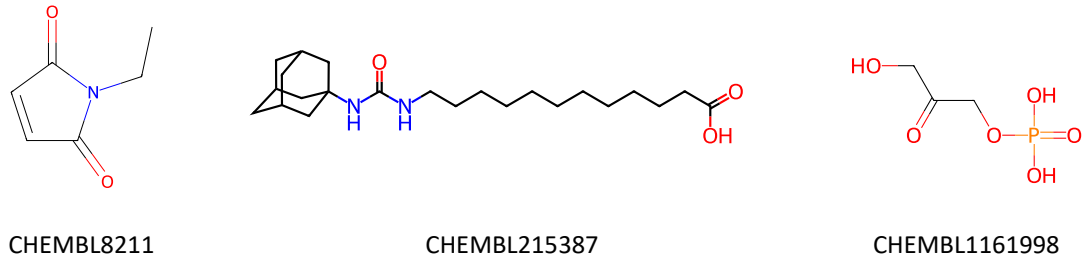


Figure 18: Compounds annotated for Q5EB52 from patent US-20120270250-A1, which share little similarity.

The lowest diversity value was obtained for the six compounds for Tachykinin-3 (Tchem, Q9UHF0) from patent US-8470816-B2, which differ only very slightly from each other (Figure 19).

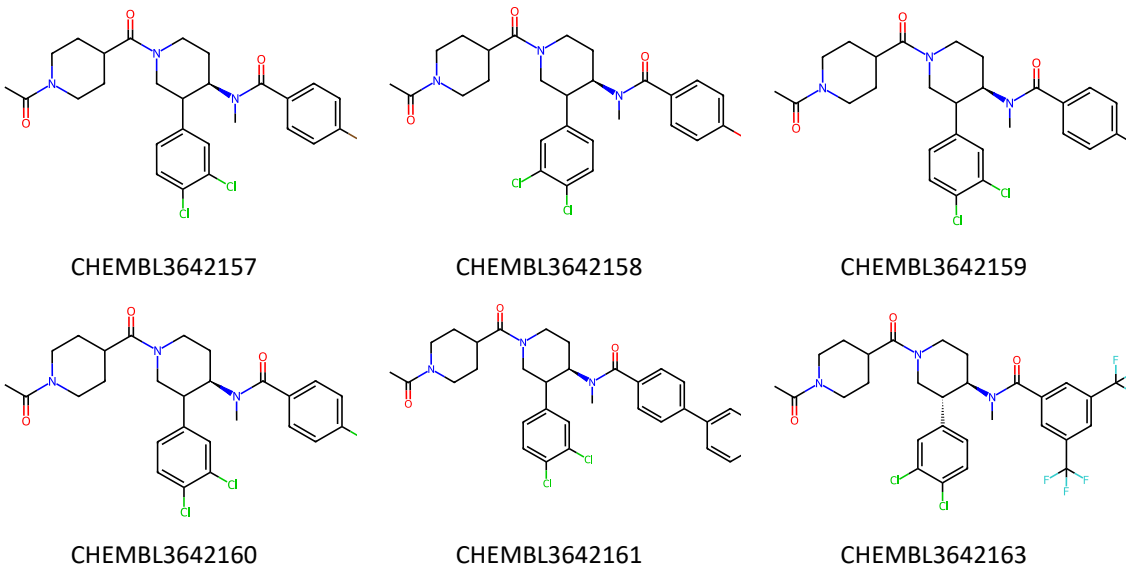


Figure 19: The six quite closely related compounds from patent US-8470816-B2 for Tachykinin-3.

Same as for fingerprint diversity, the calculation of physicochemical diversity relies on pairwise comparisons, and therefore for the 28 targets with only 1 compound the diversity was set to zero. For targets where none of the compounds had a full set of descriptors, the value was set as NaN; this is the case for 3 targets in this dataset. For those 88 targets with sufficient compounds to calculate it, the physicochemical diversity values span the range between 19.8 and 234, with the highest diversity by this measure was found for Cytochrome P450 2B6 (Tchem; P20813).

Using the quadrant thresholds established for the whole patent target dataset, the 119 targets in this dataset are distributed as follows (Figure 20): 53 of the targets have both low compound and scaffold diversity; 11 targets have high scaffold, but low compound diversity; and 18 have high compound, but low scaffold diversity. 37 targets are in the best-explored category of high compound and scaffold diversity. Among those, there are 15 Tbio targets, which best demonstrate the value gained from patent annotations. It bears repeating that for targets where no other information from scientific literature is yet present in ChEMBL 31, any data is valuable, so all the bioactivities in this dataset represent a gain in information regardless of their diversity.

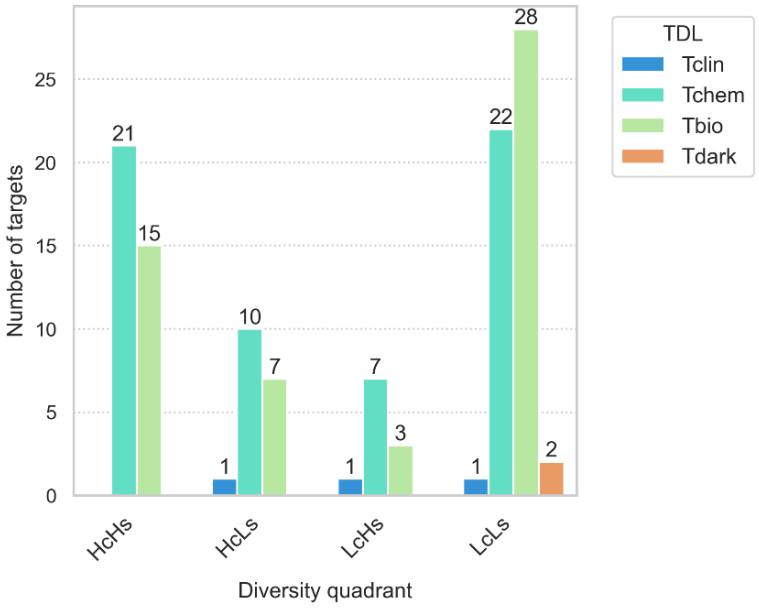


Figure 20: Distribution of targets with bioactivities only from patents from different TDLs among the diversity quadrants. HcHs, high compound and scaffold diversity; HcLs, high compound, but low scaffold diversity; LcHs, low compound, but high scaffold diversity; LcLs, low compound and scaffold diversity.

3.4 Novelty from patents

In this chapter, to assess the contribution of patent annotations to the expansion of chemical space for each target, patent-exclusive compounds were compared to those found only in non-patent sources, or both patents and non-patent literature (hereafter referred to simply as “non-exclusive”, regardless of if they appear only in non-patent sources or are shared by both). As schematically depicted in Figure 3, 1,229 of the total 1,370 human targets from patents had at least one compound annotation only found in patents. For another 119 targets, all annotations came from patents, with none available from non-patent sources. Thus, for only 22 targets no novel information at all was gained from patent annotations.

Figure 21 shows the distribution of patent-exclusive compounds among target development levels. In total, 463 targets had between one and ten patent-exclusive compounds, and 387 more targets had up to 100 patent-exclusive compounds. 46 targets even had more than a thousand. Unsurprisingly, these fall mostly into the Tclin (25 targets) or Tchem (20 targets) categories, though also one Tbio target is among them: Casein kinase II subunit beta (UniProt: P67870), with 1,176 patent-exclusive and 280 non-exclusive compounds. However, this is an exception: 75% of Tbio targets list fewer than 23 novel compounds from patents. Only four Tdark targets have patent-exclusive compounds, with a maximum count of twelve.

The highest number of novel compounds was found for the Tclin target Tyrosine-protein kinase JAK2 (UniProt: O60674), for which 4,248 patent-exclusive compounds were found; they were annotated from 44 different patents. In comparison, the non-patent literature dataset lists 5,434 molecules.

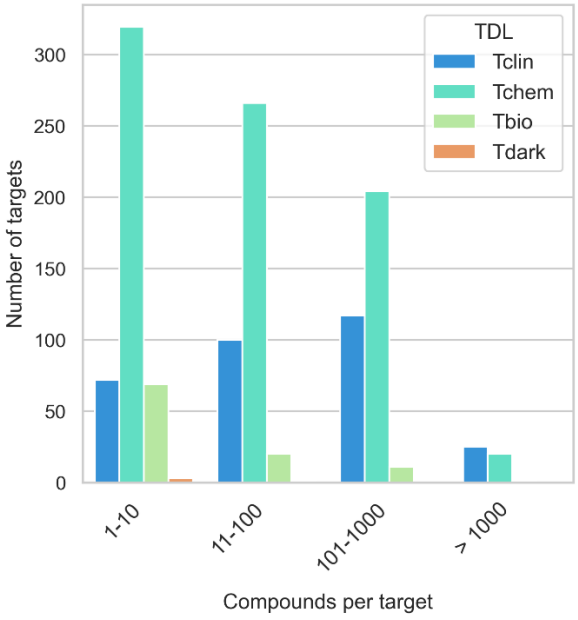


Figure 21: Number of patent-exclusive compounds, shown separately for each TDL.

However, a patent-exclusive compound may only differ slightly from its non-exclusive counterparts, therefore simple counts of exclusive molecules do not sufficiently describe the novelty of patent compounds. A similar strategy as for the diversity calculations in the last chapters was employed for a better appraisal. Scaffold novelty was evaluated via the exclusivity of *scaffolds*, which due to the nature of the Bemis-Murcko scaffold generation would not be affected by changes of, for example, a small functional group. For compound (fingerprint) novelty and physicochemical novelty, the values for each target were calculated pairwise between compounds of the patent-exclusive and the non-exclusive dataset.

Using the fraction of patent-exclusive scaffolds as replacement for the F_{50}, consensus "novelty" plots can be drawn in a similar fashion to those showing target diversity in the previous sections.

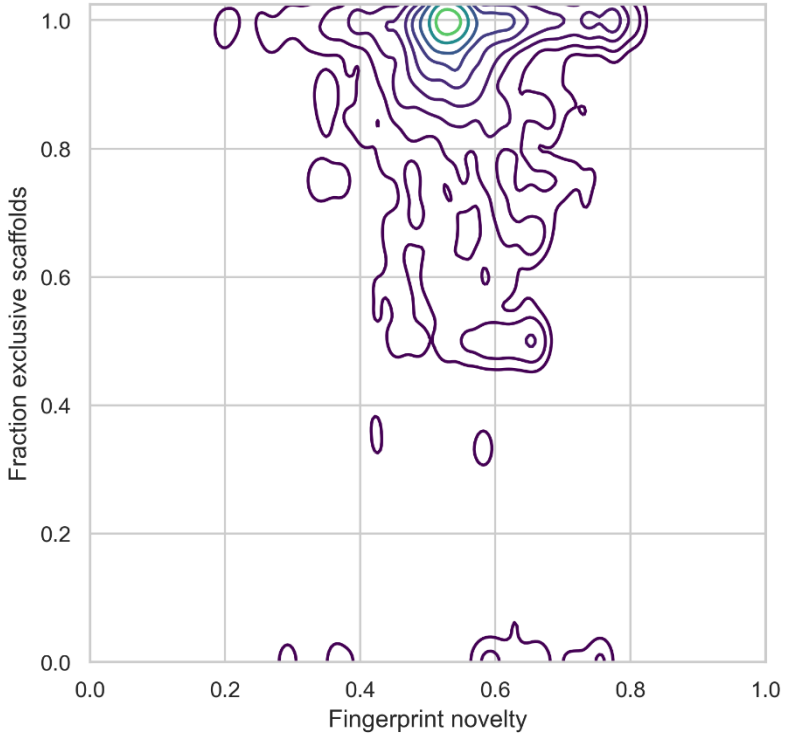


Figure 22: Consensus novelty plot for all 1,229 targets with patent-exclusive compounds.

As can be seen in the plot in Figure 22, for most targets their patent-exclusive compounds share a very small fraction of their scaffolds with those described in non-patent literature. They do not necessarily have a high fingerprint novelty as well; indeed, the scaffolds may still be quite similar to each other, as scaffold similarity is not being considered in the determination of exclusive scaffolds – much as for exclusive compound counts, only if a scaffold is identical to that of a non-exclusive compound, the scaffold is not deemed patent-exclusive.

For 629 targets, none of their patent-exclusive compounds shared a scaffold with any other compound annotated to the same target (a fraction of exclusive scaffolds equal to 1). While many targets only had a handful of exclusive compounds (344 targets had fewer than 10), others had

several hundreds, such as Casein kinase II subunit beta, already mentioned above, or G-protein coupled receptor 6 (P46095), with 907 exclusive compounds and 216 scaffolds not found in non-patent literature for this target. The stark discrepancy between patent and non-patent compounds for this target was also noted in other recent work about patent annotation in ChEMBL[17].

Only 27 targets had not a single patent-exclusive scaffold, but that was also due to a low number of patent-exclusive compounds (no more than three per target). 573 targets thus had a certain number of both shared and exclusive scaffolds.

For the pairwise comparisons in the calculation of compound and physicochemical novelty, if the mean inter-dataset diversity of a target is low, the patent-exclusive compounds share much of their features with those described in other sources, and thus add little novel knowledge.

Many of the targets with the highest compound novelty all had few exclusive compounds (below 20 for the first nine, 22 compounds for the tenth-ranked target). This highlights that this calculation may be biased if there are only few compounds in the patent dataset, especially if they are quite different from the dozens or hundreds in the non-patent dataset – the mean of the many comparisons ends up being rather high.

When excluding targets with fewer than 20 patent-exclusive compounds to circumvent this bias, the target with the highest compound novelty from patents is Pendrin (Tbio; O43511), with a value of 0.781. In ChEMBL 31, this target has 22 patent-exclusive compounds (annotated from one patent, WO-2017147523-A1) and 21 from two scientific articles, but there is little overlap in scaffolds. Interestingly, the target with the least novel compounds by this measure (Ubiquitin-like modifier-activating enzyme ATG7; Tchem; O95352) has a high fraction of patent-exclusive scaffolds among its 192 compounds, with 46 of 51 scaffolds not found in other sources, but they share a highly conserved common core.

The novelty in terms of physicochemical diversity between compounds of the two datasets could not be computed for 20 of the 1,229 targets as for them all compounds from either one or the other dataset were missing descriptors in ChEMBL. Among the other targets, the highest physicochemical novelty was found for targets with very few compounds: The five highest-ranked targets by that value all have fewer than four compounds; but importantly they have at least one patent-exclusive scaffold that presumably drives some of the differences in physicochemical descriptors.

Excluding such targets with few exclusive compounds, the highest-ranked target for physicochemical novelty is the mitochondrial Serine hydroxymethyltransferase (Tchem, UniProt: P34897). Its 235 patent-exclusive compounds share 184 unique scaffolds, none of which are found from other sources in ChEMBL 31 for that target - indeed, there is only one non-patent compound

annotated for this target. Clearly, this discrepancy contributes to the high mean pairwise Euclidian distance between descriptors of compounds from both datasets – if there is only one compound in one of the datasets, this value mostly reflects the diversity of the other dataset. (Compound novelty is not affected in quite the same way due to the binary nature of fingerprints, compared to the Euclidian distance used for the calculation of physicochemical novelty.)

For the second-highest ranked target, Glucagon-like peptide 1 receptor (Tclin, P43220), there are 105,859 compounds annotated in ChEMBL from non-patent sources, and 246 patent-exclusive ones: a ratio of 430:1. It is likely that the high diversity/novelty is due to the same reasons mentioned above.

3.4.1 Novelty by target family

As mentioned above, there are many targets whose patent-exclusive compounds do not share any scaffolds with those published in scientific literature. Table 17 lists the number of targets from each target family for which this is the case, while Figure 23 provides a visualization of each family's value distribution. The highest fraction of targets with only patent-exclusive scaffolds is found among kinases (two thirds of which share no scaffold with non-exclusive compounds), followed by GPCRs and ion channels. However, many of these targets also have relatively few patent-exclusive compounds. Thus, while the novelty by this measure may be high, the depth of the data may not always be.

Table 17: Number of targets from each family whose patent-exclusive compounds also have patent-exclusive scaffolds.

IDG Family	Number of targets	Number of targets with only patent-exclusive scaffolds	Percentage of targets with only patent-exclusive scaffolds
Kinase	461	309	67.0
GPCR	167	82	49.1
Ion Channel	66	32	48.5
Transcription Factor	13	6	46.2
Other	132	58	43.9
Enzyme	304	114	37.5
Nuclear Receptor	18	6	33.3
Epigenetic	37	12	32.4
Transporter	28	9	32.1
TF-Epigenetic	2	0	0

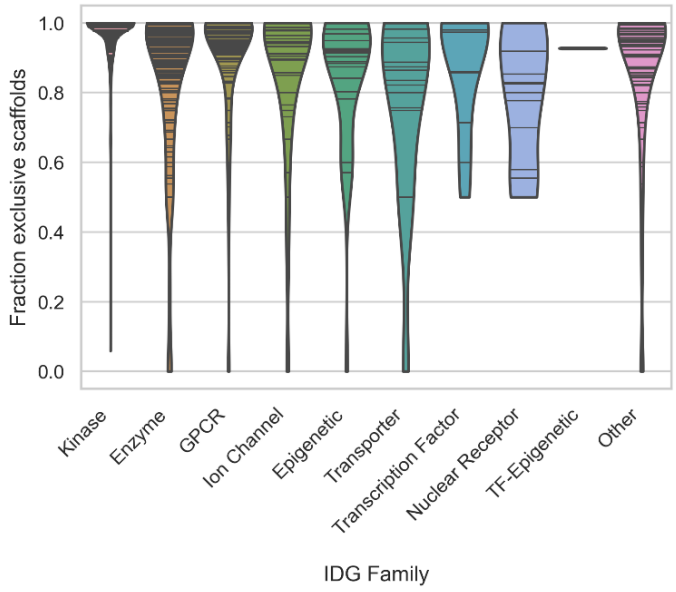


Figure 23: Distribution of target scaffold novelty values among the different families. Lines within each violin designate separate targets; a wider violin body indicates more targets in that area, however the maximum width does not represent the same number of targets in each family, as there are large discrepancies in numbers.

For most target families, the fingerprint novelty of individual targets covered a broad range, with some targets having patent-exclusive compounds that are relatively similar to compounds published elsewhere, and others with high novelty of patent compounds. The highest values in most families were around 0.8 (Figure 24). Among the larger families, only kinases show considerably less variability; even compounds for kinases with high scaffold novelty from patents appear to share many substructures with compounds described elsewhere.

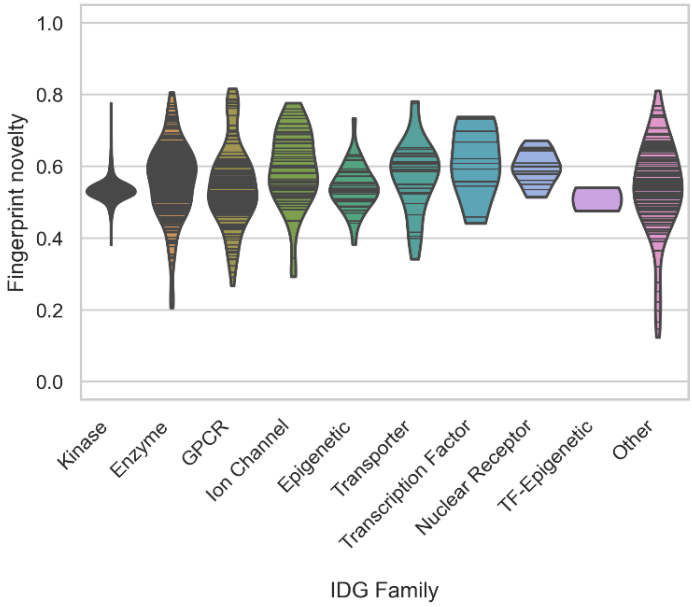


Figure 24: Distribution of target fingerprint novelty values among the different families.

Regarding novelty in terms of physicochemical diversity, while there are differences in the number of greatly diverging values, the overall distribution is similar across families (Figure 25). The extreme values are likely due to an imbalance of the numbers of compounds from each dataset used for the comparisons, as mentioned above.

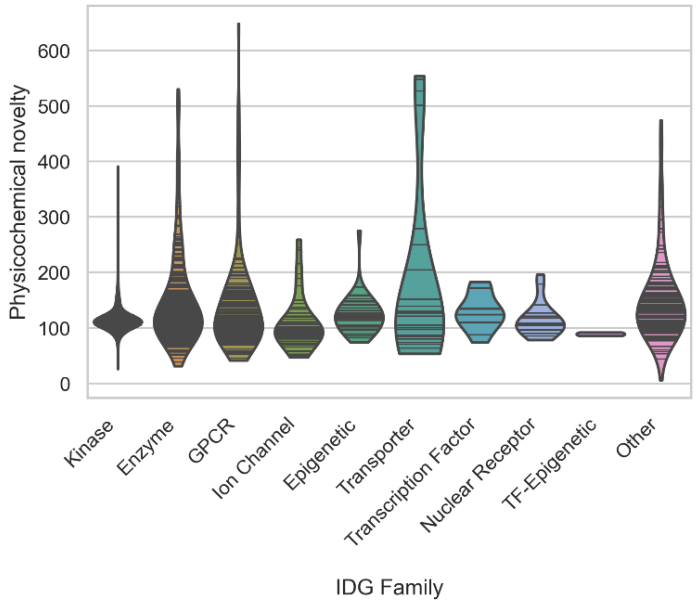


Figure 25: Distribution of physicochemical novelty values among the different families.

Examining the consensus novelty plots for individual families, it is striking that kinases, GPCRs and ion channels, the main focus areas of the IDG, have much more uniformly high scaffold novelties than the other two groups with roughly equal numbers of targets in the dataset, enzymes and “other” targets (Figure 26).

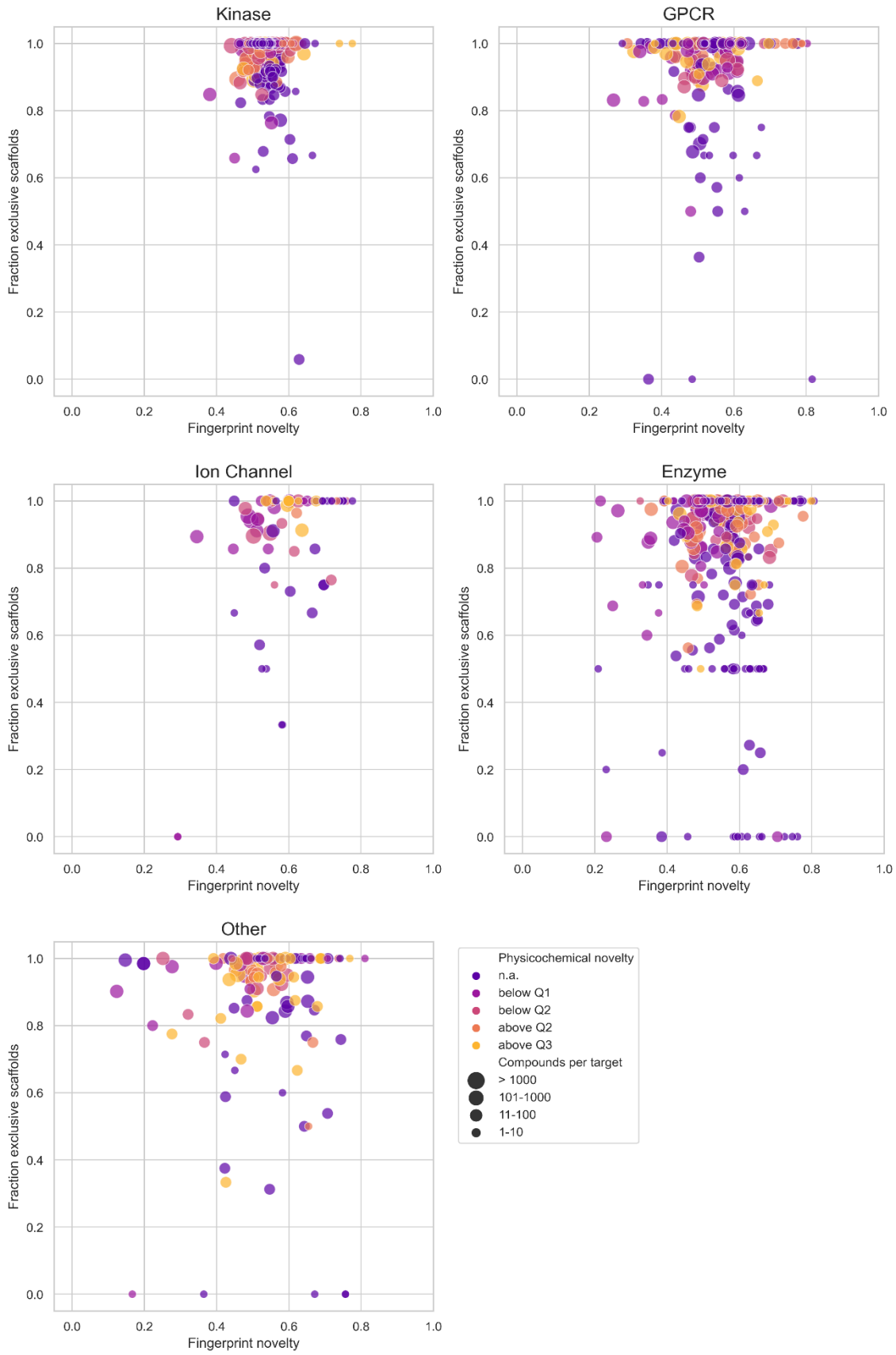


Figure 26: Consensus novelty plots for the larger families: Kinases, GPCRs, ion channels, enzymes and "other".

For further analysis of the patent contributions in terms of novelty of compounds, targets were grouped into novelty quadrants, analogous to the process in the previous chapters for diversity. As a threshold for the scaffold novelty, expressed as the fraction of scaffolds exclusive to patents, a value of 0.8 was chosen. Most targets thereby fall into the high scaffold novelty quadrants, as many targets with few compounds have a value of 1; however, setting the threshold higher may exclude targets with a decent number of compounds but also a few shared scaffolds. The threshold for the novelty in terms of fingerprint diversity was set separately for each family as the third quartile (Q3), encompassing the highest 25% of values. Physicochemical diversity is not included in the assignment of a novelty quadrant, as it is strongly biased by the number of compounds in both datasets, and it is hard to determine a useful threshold for a non-normalized value.

The classification into novelty quadrants resulted in 246 targets in the highest category (high scaffold and compound novelty), 63 targets with high compound, but low scaffold novelty, 800 targets with low compound, but high scaffold novelty, and 120 targets with overall low novelty. Table 18 and Figure 27 show the distribution in each family. It should be kept in mind that due to the rather lax threshold for scaffold novelty, most targets are found in the HcHs and LcHs quadrants. Of the larger families, the least novel compounds were found for ion channels and enzymes, with 18 and 17% of targets in the lowest novelty quadrant, respectively.

Table 18: Numbers of targets from each family in the different novelty quadrants. HcHs, high compound and scaffold novelty; HcLs, high compound, but low scaffold novelty; LcHs, low compound, but high scaffold novelty; LcLs, low compound and scaffold novelty.

IDG Family	Number of targets	HcHs	HcLs	LcHs	LcLs
Kinase	461	109	6	342	4
Enzyme	304	47	29	175	53
GPCR	167	37	5	107	18
Ion Channel	66	14	3	37	12
Epigenetic	37	6	3	23	5
Transporter	28	3	4	14	7
Transcription Factor	13	1	2	9	1
Nuclear receptor	18	3	2	7	6
TF-Epigenetic	2	1		1	
Other	132	24	9	85	14

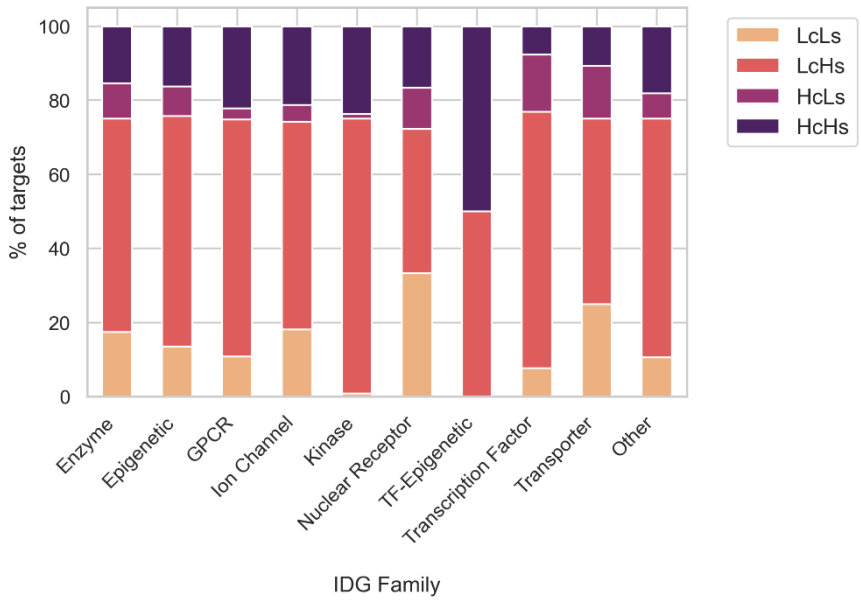


Figure 27: Novelty distribution of targets in each family. The large proportion of targets in the high scaffold novelty quadrants is due to a rather low cutoff of 0.8. HcHs, high compound and scaffold novelty; HcLs, high compound, but low scaffold novelty; LcHs, low compound, but high scaffold novelty; LcLs, low compound and scaffold novelty.

Examining the quadrants for targets of the different development levels, a greater fraction of understudied targets from the Tbio and Tdark levels is in the highest compound and scaffold novelty categories (Table 19 and Figure 28). While this is not unexpected – after all, these are targets where there is little data, so additional information has a greater impact than for a well-studied target – it is a confirmation of the value of patent annotations for expanding the chemical space. Interestingly, around 12% of Tclin targets also had high novelty patent-exclusive compounds, as well as 20% of Tchem targets.

Table 19: Numbers of targets from each TDL across novelty quadrants. HcHs, high compound and scaffold novelty; HcLs, high compound, but low scaffold novelty; LcHs, low compound, but high scaffold novelty; LcLs, low compound and scaffold novelty.

TDL	total	HcHs	HcLs	LcHs	LcLs
Tclin	314	38	16	214	46
Tchem	809	166	43	533	67
Tbio	101	38	4	52	7
Tdark	4	3		1	

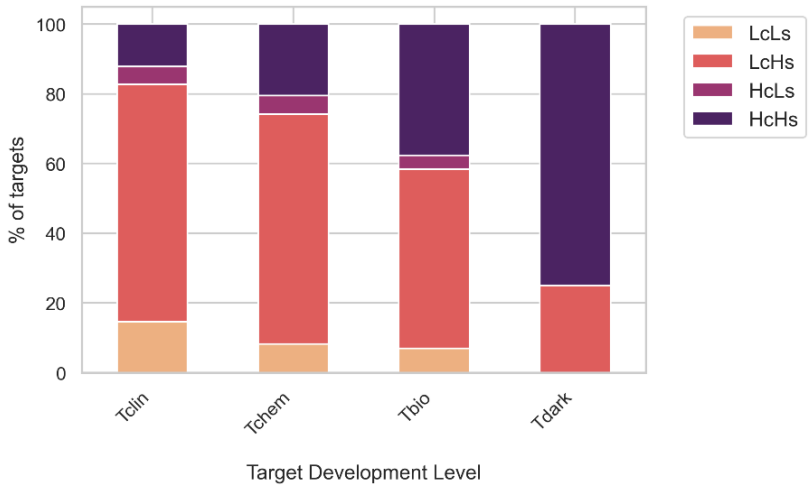


Figure 28: Distribution of targets from each TDL across the novelty quadrants. HcHs, high compound and scaffold novelty; HcLs, high compound, but low scaffold novelty; LcHs, low compound, but high scaffold novelty; LcLs, low compound and scaffold novelty.

124 targets from the IDG priority list have patent-exclusive compounds in ChEMBL; 110 of them are Tchem, 11 Tbio and 3 Tdark (Table 20). Kinases make up the largest number with 94 targets, followed by 29 GPCRs and one ion channel. As shown in Figure 29, their target-wise novelty tends to be rather high: Only three targets are in the lowest-novelty section, while 40 are high in both compound and scaffold novelty for their respective family, and for the remaining 81, scaffold novelty is high while compound novelty is not.

Table 20: Tbio and Tdark targets from the IDG priority list with patent-exclusive compounds in ChEMBL, all of which happen to be kinases.

UniProt accession	Name	IDG Family	TDL
Q96S44	EKC/KEOPS complex subunit TP53RK	Kinase	Tbio
Q6P0Q8	Microtubule-associated serine/threonine-protein kinase 2	Kinase	Tbio
Q5TCX8	Mitogen-activated protein kinase kinase kinase 21	Kinase	Tbio
Q5TCY1	Tau-tubulin kinase 1	Kinase	Tbio
Q6IQ55	Tau-tubulin kinase 2	Kinase	Tbio
Q86Y07	Serine/threonine-protein kinase VRK2	Kinase	Tbio
Q99755	Phosphatidylinositol 4-phosphate 5-kinase type-1 alpha	Kinase	Tbio
Q8NI60	Atypical kinase COQ8A, mitochondrial	Kinase	Tbio
P31152	Mitogen-activated protein kinase 4	Kinase	Tbio
Q9NRP7	Serine/threonine-protein kinase 36	Kinase	Tbio
Q86UX6	Serine/threonine-protein kinase 32C	Kinase	Tbio
Q96S38	Ribosomal protein S6 kinase delta-1	Kinase	Tdark
Q58A45	PAN2-PAN3 deadenylation complex subunit PAN3	Kinase	Tdark
Q8N752	Casein kinase I isoform alpha-like	Kinase	Tdark

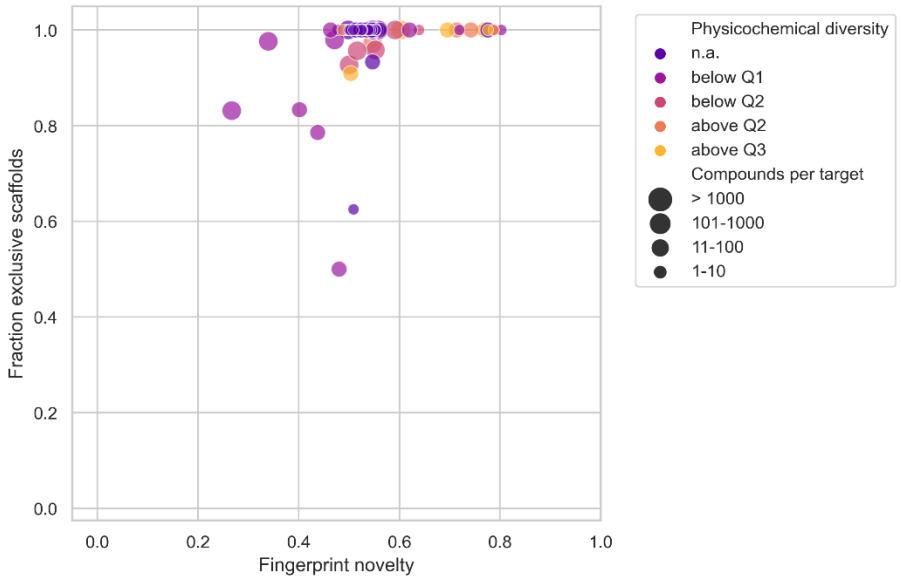


Figure 29: Consensus novelty plot of IDG priority list targets in the dataset.

Although the total target numbers may not be enough to draw a firm conclusion, it appears that IDG priority list targets are enriched in the higher novelty quadrants compared to targets from the same families but not on the priority list (Table 21). If examined in more detail, this seems mainly to be due to GPCRs, as only one priority ion channel is on the list (it is in the HcHs novelty quadrant though), and for kinases the distribution is largely equal to their non-priority list counterparts (Table 22).

Table 21: Comparison of IDG priority and non-priority targets in respect to their novelty classification. HcHs, high compound and scaffold novelty; HcLs, high compound, but low scaffold novelty; LcHs, low compound, but high scaffold novelty; LcLs, low compound and scaffold novelty.

	total	HcHs	HcLs	LcHs	LcLs
IDG priority targets	124	40 (32%)		81 (65%)	3 (2%)
Non-IDG priority kinases, GPCRs, ion channels	570	120 (21%)	14 (3%)	405 (71%)	31 (5%)
All kinases, GPCRs, ion channels	694	160 (23%)	14 (2%)	486 (70%)	34 (5%)

Table 22: Detailed novelty comparison of kinases, GPCRs and ion channels included or not included on the IDG priority list. HcHs, high compound and scaffold novelty; HcLs, high compound, but low scaffold novelty; LcHs, low compound, but high scaffold novelty; LcLs, low compound and scaffold novelty.

	IDG priority kinases	Other kinases	IDG priority GPCRs	Other GPCRs	IDG Priority Ion Channels	Other ion channels
Total	94	367	29	120	1	65
HcHs	22 (23%)	87 (24%)	17 (59%)	2 (2%)	1 (100%)	13 (20%)
HcLs		6 (2%)		5 (4%)		3 (5%)
LcHs	71 (76%)	271 (74%)	10 (35%)	97 (80%)		37 (57%)
LcLs	1 (1%)	3 (1%)	2 (7%)	16 (13%)		12 (19%)

3.5 Novelty of active compounds from patents

The previous chapter discussed the novelty of patent-exclusive compounds in general, and while there is use in compound annotations regardless of the activity, there naturally is great interest in actives. For the purposes of this analysis, the family-wise thresholds from IDG, detailed in Table 5 in the methods section, were used to designate compounds as active or inactive. As mentioned before, this does not take into account compounds that are described in the text as active; only bioactivity data that was converted to a pChEMBL value is considered here.

The patents included in ChEMBL31 yielded patent-exclusive actives for 725 human targets for which also non-exclusive actives were found (see Figure 3 for a schematic of the datasets in this work). For another 414 targets, actives were only found in non-patent literature; none of the patent bioactivity annotations met the respective cutoff. For 48 targets, patents were the only source of actives. In addition, actives were found for 54 targets without any non-patent bioactivity annotations. The diversity of active compounds for these 104 targets, which can be regarded as very novel since they do not occur in non-patent literature, is discussed separately in section 3.6.

For 677 targets, therefore, actives from both types of sources were found, and can be compared to estimate the novelty of actives from patents. The majority (407 targets) were Tchem, while 256 targets were Tclin and 13 Tbio. Notably, the annotations for Tbio targets all had “protein complex” as target type, which prevents them from appearing in Pharos and progression of the targets to Tchem.

While 56 targets had only a single patent-exclusive active compound listed, others had several thousand; the distribution in each family is shown in Figure 30. The highest number of patent-exclusive actives was for the Tclin target PDE10A (cAMP and cAMP-inhibited cGMP 3',5'-cyclic phosphodiesterase 10A; UniProt: Q9Y233) with 2,996 active compounds only found in patents, versus 1,167 found in non-patent literature. 14 targets only had one active compound from non-patent literature listed in ChEMBL; for ten of those, there were more actives found exclusively in patents, with up to 402 patent-exclusive active compounds for the Interleukin-23 receptor (Tchem, Q5VWK5).

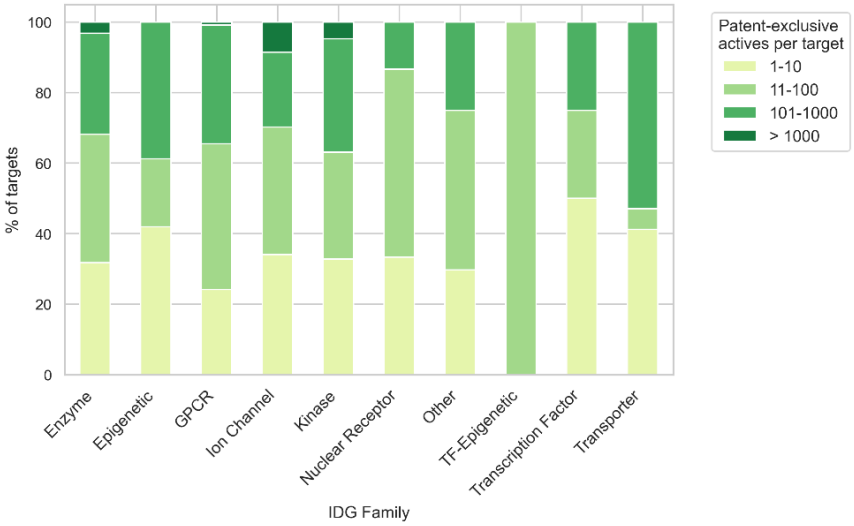


Figure 30: Distribution of number of actives per target in each family.

501 targets had more actives from non-patent sources than patent-exclusive active compounds, but the opposite was true for 168 targets (8 targets had the same number of patent-exclusive actives and actives from non-patent literature).

Figure 31 shows a contour plot of the distribution of target novelties regarding scaffolds and compounds. Clearly, the scaffold novelty of patent-exclusive actives is very high: For 281 of the 677 targets, all of their patent-exclusive actives' scaffolds were novel. For around half of them (145) the number of actives was below 10, but there are also targets with hundreds of actives, all of which have novel scaffolds. Among them, the highest number of patent-exclusive scaffolds was found for Dipeptidyl peptidase 1 (Tchem, UniProt: P53634), whose 574 patent-exclusive actives share 245 scaffolds not found in non-patent literature.

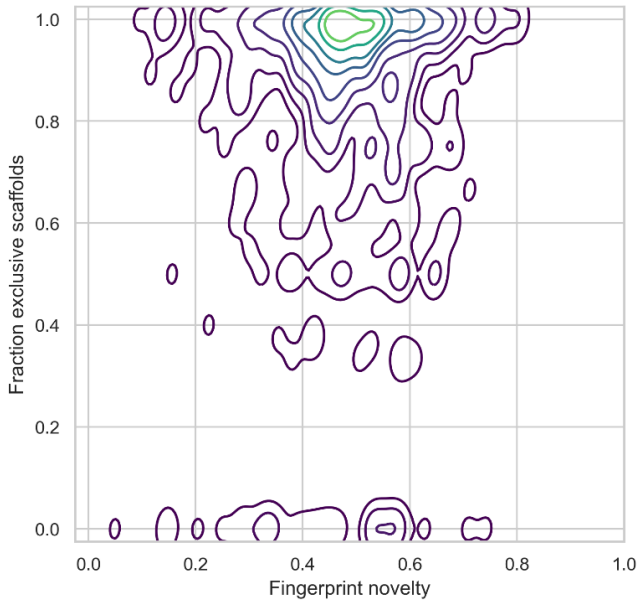


Figure 31: Consensus novelty plot for patent-exclusive actives per target.

As in the previous chapter, compound novelty is assessed here by the mean pairwise fingerprint diversity between patent-exclusive and non-exclusive actives for each target. Compound novelty is more evenly distributed among the whole dataset than scaffold novelty. The highest compound novelty value was found for the histone lysine demethylase PHF8 (Tchem, Q9UPP1), which has 49 exclusive actives with 18 unique scaffolds, but only two actives from non-patent sources. This ratio is more balanced for the target with the second-highest compound novelty regarding actives when excluding targets with fewer than 20 compounds, Prostaglandin E synthase 2 (Tchem, Q9H7Z7), which has 113 patent-exclusive and 20 non-exclusive compounds, with no overlap in scaffolds between them.

For 26 targets, the physicochemical novelty could not be calculated due to the same reasons outlined in the last chapter. The same bias concerning large differences in compound numbers between the datasets mentioned above was encountered here: 12 targets had more than 20-fold more patent-exclusive actives than those from non-patent literature, while the reverse was the case for 164 targets.

Among the remaining 485 targets, the one with the highest physicochemical novelty for its actives was Fibroblast growth factor 2 (Tchem, P09038), which has 3 patent-exclusive actives whose scaffolds do not overlap with the 57 non-exclusive actives. While most of the non-exclusive compounds are polysulfated glycosides, those annotated from the single patent (US-8933099-B2) are pyridine derivatives.

3.5.1 Novelty of actives by target family

Figure 32 shows the consensus novelty plots of the IDG priority list families: Kinases, GPCRs and ion channels. The overall distribution of targets within them is quite similar, with most patent-exclusive actives for each target also possessing exclusive scaffolds, but a varying degree of compound novelty. There are very few targets whose patent-exclusive actives share a large fraction of their scaffolds with actives published elsewhere. In comparison to the other two families, it appears GPCRs have a rather high physicochemical novelty, though as discussed before, that value needs to be interpreted cautiously.

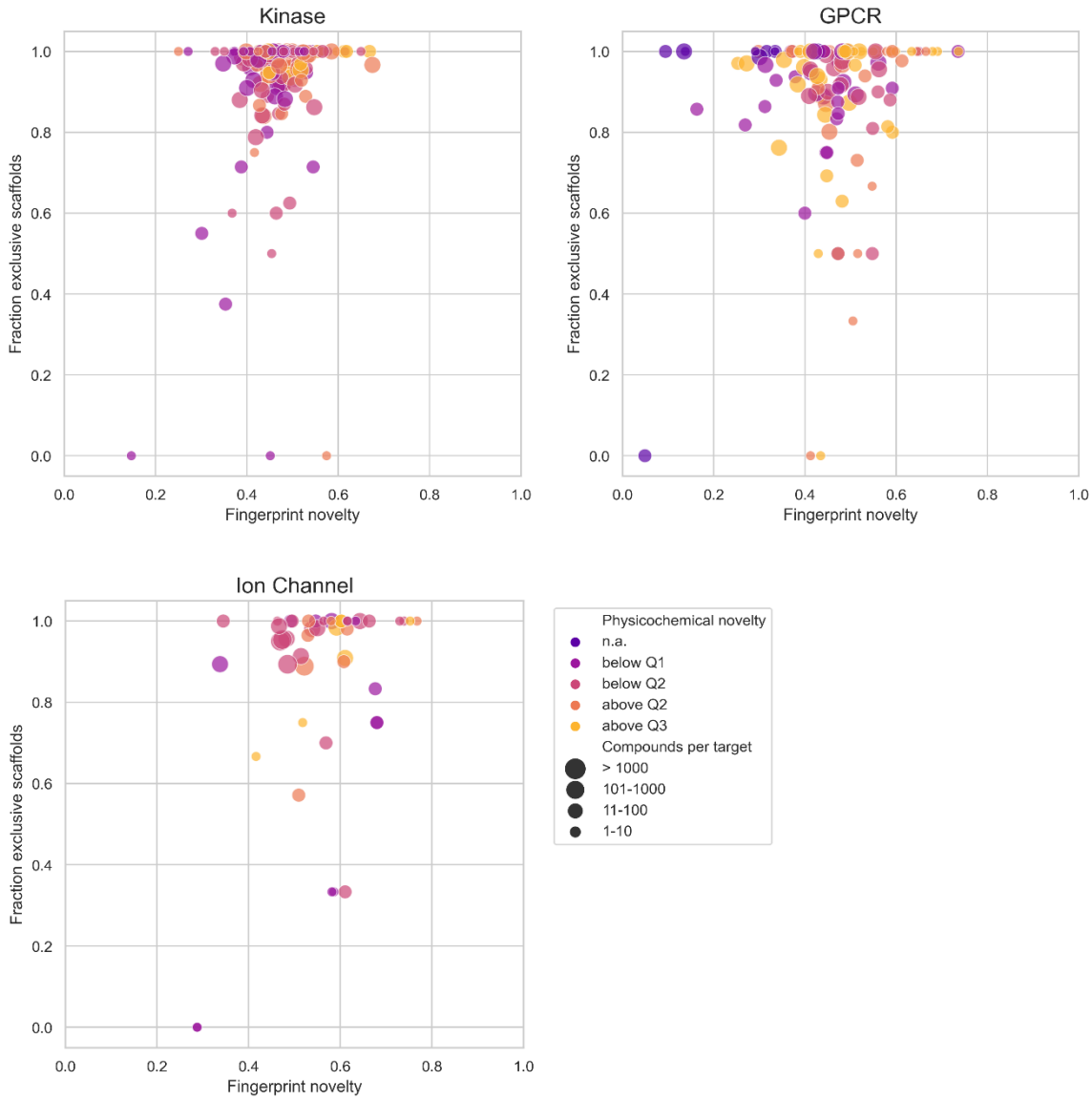


Figure 32: Consensus novelty plots for actives from kinase, GPCR and ion channel targets.

For the other two large target categories, enzymes and "other", targets are distributed somewhat more widely in terms of scaffold novelty (Figure 33). They also cover a larger range of compound novelty than the IDG priority families; some target's patent-exclusive actives appear to be quite similar to those published in non-patent literature. For the other families, only few targets of each had patent-exclusive actives, rendering consensus novelty plots somewhat futile; thus they are not shown.

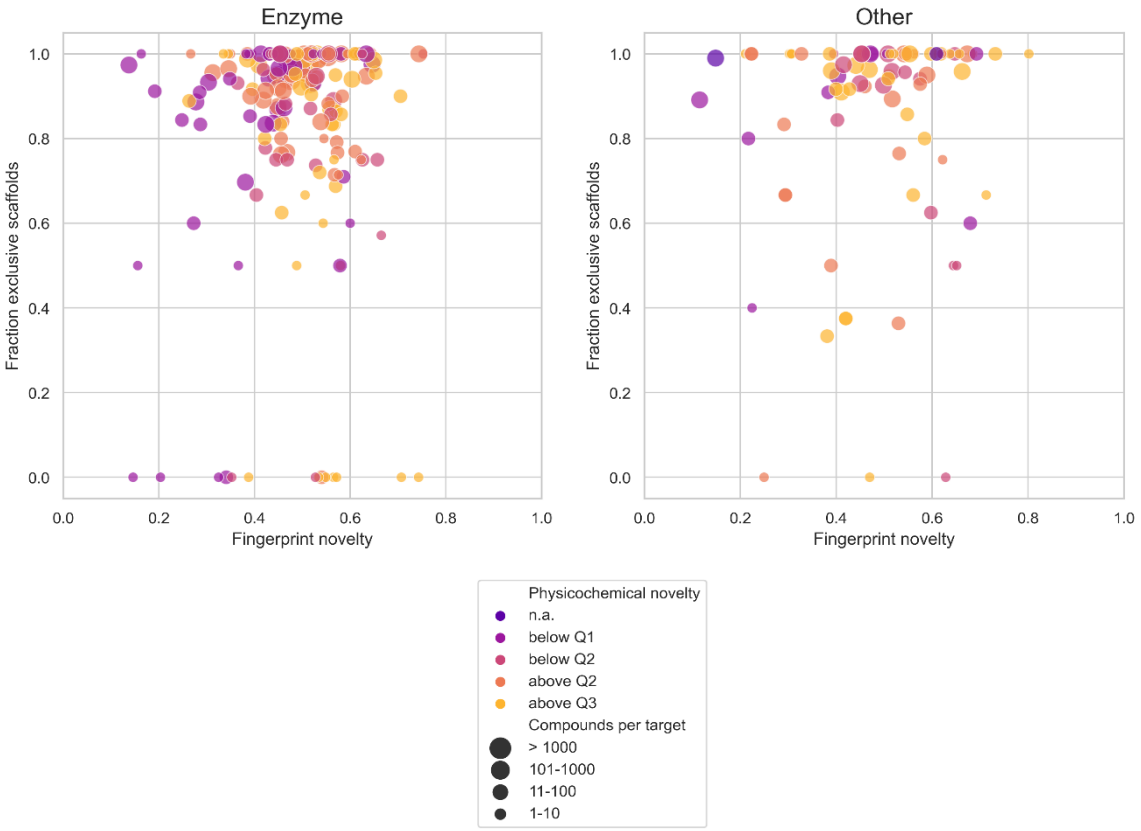


Figure 33: Consensus novelty plots of patent-exclusive actives for enzyme and "other" targets.

Comparing the fraction of patent-exclusive scaffolds for all target families (Table 23), it is quite uniformly very high, with only some of the smaller groups showing median values below 0.9. Table 23 shows these medians as well as the percentage of targets with only patent-exclusive scaffolds, which also is relatively similar (between 40 and 50%) for those families with a decent number of targets. Only enzymes have a lower percentage of around one third.

Table 23: Novelty descriptors for each family.

IDG Family	total	Median fraction of exclusive scaffolds	Targets with only exclusive scaffolds
Kinase	168	0.995	83 (49%)
Enzyme	192	0.948	65 (34%)
GPCR	116	0.967	48 (41%)
Ion Channel	47	0.981	20 (43%)
Epigenetic	31	0.977	14 (45%)
Transporter	17	0.885	4 (24%)
Transcription Factor	4	1	4 (100%)
Nuclear receptor	15	0.818	5 (33%)
TF-Epigenetic	2	0.84	0
Other	84	0.959	37 (44%)

To assess the combined scaffold and compound novelty of actives, targets were grouped into novelty quadrants according to the same criteria used in the previous chapter: For scaffold novelty, a threshold of 0.8 was used again in order to not exclude targets with a few overlapping scaffolds

that are otherwise high in novelty; and for compound novelty, the third quartile for each target family was used.

Due to the relatively liberal scaffold novelty cutoff, most targets were in the higher quadrants for this parameter again: 134 targets in the highest novelty quadrant (HcHs) and 414 in the low compound, high scaffold novelty (LcHs) quadrant. 37 targets had high compound, but low scaffold novelty, and the remaining 92 targets were low in novelty according to both parameters.

Among the larger families, kinases have the highest percentage of targets in the HcHs quadrant (24%), followed by GPCRs and ion channels (see Table 24 and Figure 34). In contrast, enzymes, the most numerous family among the 677 targets, has the comparatively low percentage of 15% high compound and scaffold novelty targets. A similarly low portion is reached by those targets collected in the "other" category.

Table 24: Numbers of targets from each family in the different novelty quadrants. HcHs, high compound and scaffold novelty; HcLs, high compound, but low scaffold novelty; LcHs, low compound, but high scaffold novelty; LcLs, low compound and scaffold novelty.

IDG Family	Number of targets	HcHs	HcLs	LcHs	LcLs
Kinase	168	40	2	114	12
Enzyme	192	29	19	115	29
GPCR	116	26	3	72	15
Ion Channel	47	10	2	26	9
Epigenetic	31	8		20	3
Transporter	17	2	2	8	5
Transcription Factor	4	1		3	
Nuclear receptor	15	2	2	6	5
TF-Epigenetic	2	1		1	
Other	84	14	7	49	14

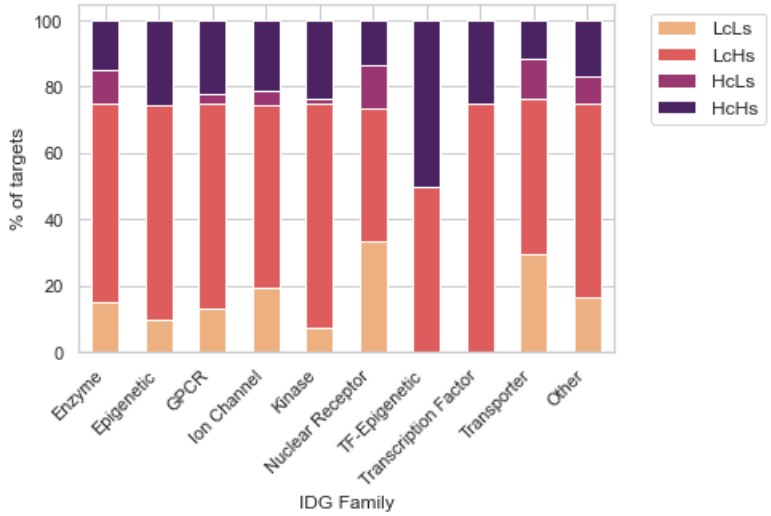


Figure 34: Distribution of targets among novelty quadrants. HcHs, high compound and scaffold novelty; HcLs, high compound, but low scaffold novelty; LcHs, low compound, but high scaffold novelty; LcLs, low compound and scaffold novelty.

22 targets from the IDG high priority list are found in this dataset: 11 kinases, 19 GPCRs and one ion channel. Their separate consensus novelty plot is shown in Figure 35. All of them have high scaffold novelty; the ion channel, four kinases and three GPCRs also were classified as having high compound novelty. Considering that these are understudied targets by definition, it is not surprising that they show high novelty among their actives, but it is nevertheless a signal that the patent extractions are proving valuable and new information.

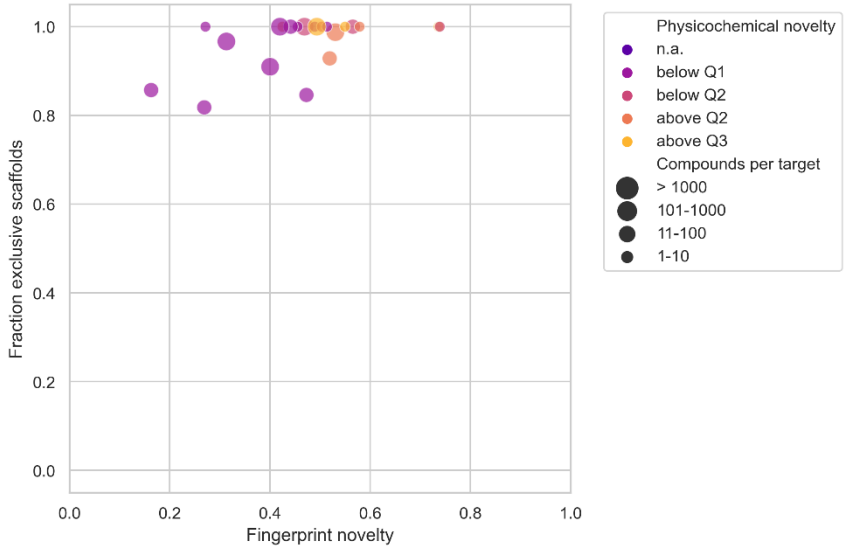


Figure 35: Consensus novelty plot for patent-exclusive actives per IDG priority target.

Aside from considering the expansion of chemical space for a target with the annotation of new actives from patents, a reasonable question to ask is whether those actives differ from the ones in non-patent literature in terms of potency.

For 335 of the 677 targets, actives exclusive to patents had a higher median pChEMBL value than those occurring in non-patent literature. The reverse is true for 331 targets; 11 targets had the same median pChEMBL values among both sets of actives. Overall, there seems to be no trend towards higher or lower activities among patent-exclusive compounds, neither is that seen when comparing the target families as a whole (Figure 36). Indeed, for some families like epigenetic targets, the values of patent-exclusive actives tend to be lower, and generally there is a greater variability, seen especially clear in kinases and nuclear receptors.

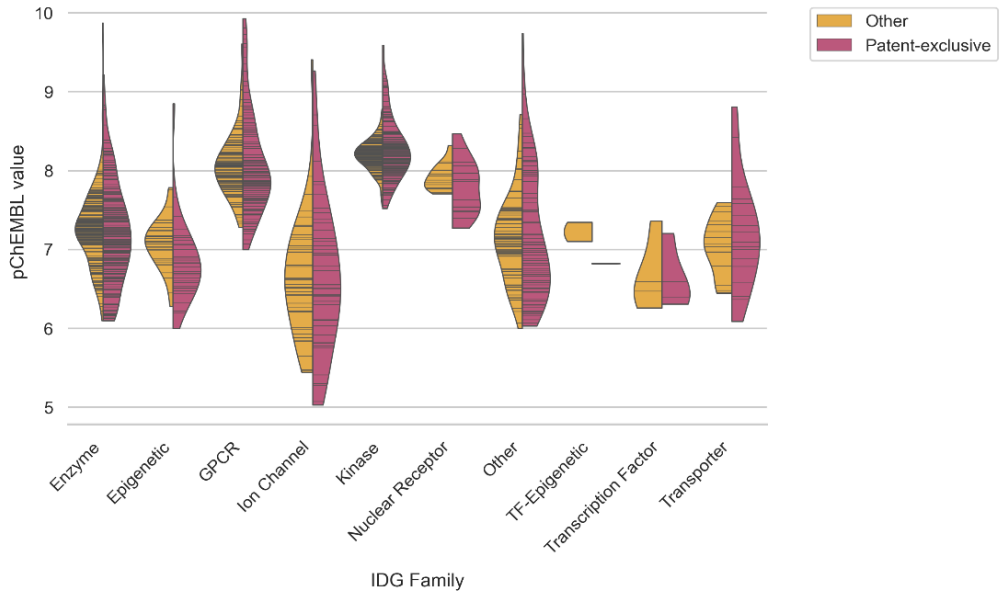


Figure 36: Distribution of target median pChEMBL values from patent-exclusive actives and other actives. Note that the threshold for actives defined by the IDG depends on the family (Kinases: 7.52; GPCRs and nuclear receptors: 7; ion channels: 5; all other families: 6).

3.6 Novelty of patent actives for targets without non-patent actives

As mentioned above, there are 104 targets for which actives have been annotated from patents, but not other sources. The majority of these, 97 targets, are Tchem, but there are also five Tclin and one Tbio target, as well as one of the two targets not yet included in Pharos (DGAT1). As can be seen from Table 25, targets from most families are present in this subset. Considering that they were very numerous in the other datasets analyzed above, it may seem surprising that there are only three kinases where actives were only annotated from patents. One explanation may be that there are compounds which are capable of inhibiting whole subgroups of the kinase family[44], leaving fewer without any actives from non-patent literature. Of course, kinases are also among the targets of highest interest in general and may simply be better explored, with few remaining which have not been targeted with actives published outside of patents.

Table 25: Targets with actives exclusively from patent annotations.

IDG Family	Number of targets
Kinase	3
Enzyme	23
GPCR	29
Ion Channel	4
Epigenetic	3
Transporter	3
Transcription Factor	5
Nuclear receptor	1
oGPCR	1
Other	31

The sole Tbio target without actives from other sources is the transcription factor Protein max (UniProt: P61244), for which one active compound with a pChEMBL value of 7.87 was found. However, it is not included in Pharos, preventing the target to be classified as Tchem, presumably as it is annotated for the protein complex of the protein with c-Myc.

Of the five Tclin targets in this subset, three were already mentioned in earlier sections and are not discussed here in detail (Sclerostin as well as Potassium voltage gated channel subfamily A member 7 and subfamily H member 3, see section 3.3). The other two are Ryanodine receptor 2 (Q92736) and Receptor activity-modifying protein 3 (Gene symbol: RAMP3; UniProt: O60896). For the former, ChEMBL lists 52 active compounds, all from the patent EP-2764867-A1 from 2014. Pharos lists another compound (VK-II-86; ChEMBL ID: CHEMBL2440787) for which no annotation to this target is found in ChEMBL (the article was annotated but "unchecked" was selected as target type, possibly as the assay targets a mutant of the target protein); the same applies for its approved drugs carvedilol (CHEMBL723; same article as VK-II-86) and hydralazine (CHEMBL276832, for which the article referenced in Pharos was not annotated.).

RAMP3 is an interesting example of a target with Tclin status but no active ligands in Pharos. This appears to be due to its actives (including the approved peptide drug pramlintide) being annotated to two receptors whose target type is "protein complex" containing RAMP3 in ChEMBL. One of them, the Amylin receptor AMY3 (consisting of CALCR/RAMP3) has 141 active compounds from one patent annotated in ChEMBL. The other target (Adrenomedullin receptor; complex of CALCRL and RAMP3), lists 8 actives from scientific literature.

The target with the most numerous actives in this dataset is the Tissue factor pathway inhibitor (Tchem, UniProt: P10646), which was already mentioned in section 3.3 as patents are generally its only source of bioactivity annotations. Of its total 2,010 compounds, 1,128 satisfy the criteria for actives in the context of this project. In contrast, 37 of the 104 targets have only a single respective active compound in ChEMBL 31, which were not found in scientific articles.

Among those 48 targets for which non-patent literature yielded bioactivity annotations, but no actives, the target with the most patent-exclusive actives (460 compounds from one patent) is the Histone-lysine N-methyltransferase SUV39H2 (Tchem, Q9H5I1); a success in patent annotation highlighted also in a recent publication[17].

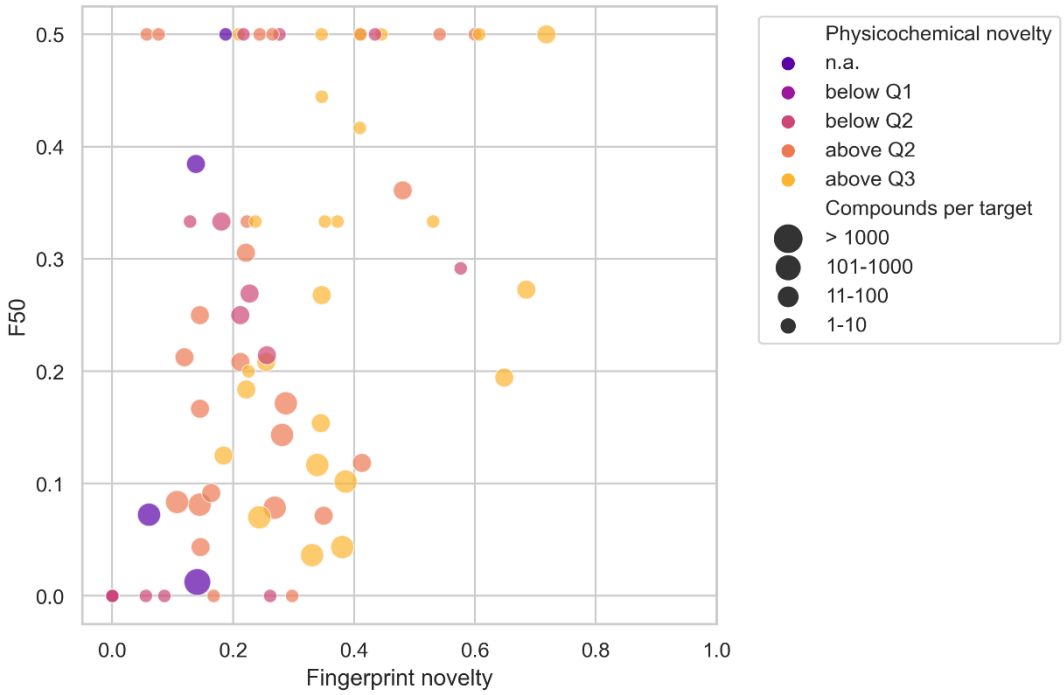


Figure 37: Consensus novelty plot for targets with no non-patent actives. Note that F_{50} is used for scaffold novelty here as all scaffolds are per definition exclusive in this dataset.

Figure 37 shows the novelty for all targets in this dataset. Most of them show relatively low scaffold diversity among their actives. Considering that in most cases (84 of the 106 targets) these compounds originate from a single patent, this is rather expected, as most patents describe relatively closely related compounds. Likewise, their fingerprint diversity tends to be quite low.

Interestingly, the target with the lowest scaffold diversity (determined by F_{50}) is the Tissue factor pathway inhibitor; the majority of its actives is covered by just four of its total 267 scaffolds. This clearly shows that the number of scaffolds associated with a target is not necessarily a good measure of its scaffold diversity. Of those targets with the maximum F_{50} value of 0.5, most have fewer than 10 actives, except for the olfactory receptor 51E2 (Tchem, Q9H255) with 18 compounds. With a fingerprint diversity of 0.718, this is the target with the most diverse actives in this dataset.

Of the 67 targets that had more than one active compound, the physicochemical diversity could not be calculated for 5 targets due to missing descriptors. Among the remaining ones, the highest diversity was found for the 16 actives of the E3 ubiquitin-protein ligase TRIM33 (Tchem, Q9UPN9), which also has one of the highest compound diversity values in this dataset.

This dataset also includes 26 targets found on the IDG priority list; 22 of them have a sole active compound, while three others have 2-6 actives. The great exception is the G-protein coupled receptor 6 (Tchem; UniProt: P46095) with 350 actives from 8 patents, but it generally lacks bioactivity data aside from patents.

3.7 Summary of results

In the previous sections, the contribution of patent annotations to the bioactivity information in ChEMBL 31 was examined in terms of diversity as well as novelty compared to data from other sources. Active compounds, which hold special interest, were analyzed separately in more depth. These sections cover the question of how much and what information was annotated from patents.

Another question at the beginning of this project was which targets would benefit most from annotating patents. Clearly, that would be targets with no information from other sources in ChEMBL (such as those covered in section 3.3) or those with no active compounds from non-patent literature (section 3.5); especially important among them are Tdark and Tbio targets. An example of successful illumination of targets through patent annotations are those 53 Tbio and two Tdark targets for which no other source of bioactivity data was found in ChEMBL. Focusing on active compounds only, mostly Tchem and Tclin targets profited from extraction of bioactivities above the IDG cutoffs from patents, but such were also found for one Tbio target.

Another group of targets of high interest are those included on the IDG priority list, of which 131 have patent annotations in ChEMBL. Some of these targets have rather high diversity among their annotated compounds (section 3.2). For the majority of them (124 targets) the compounds that were annotated from patents were not found in non-patent literature, and also share little scaffold or compound similarity (section 3.4). 48 IDG priority targets also had patent-exclusive actives, which in the case of 26 targets were the only actives found in ChEMBL.

One of the most surprising findings was that ChEMBL 31 contained patent annotations for two targets that were not yet manually reviewed in UniProt and therefore also not included in Pharos or given a TDL: DGAT1 and FML2. For the latter, ChEMBL also had entries and even one active compound from non-patent sources, but the patent-exclusive ones still were high in scaffold and compound novelty. In contrast, the compounds and actives for DGAT1 were annotated solely from one patent.

Before discussing targets from the different development levels in more detail, it is worth reiterating here that the analyses in this project did not consider temporal relationships. Considering that ChEMBL is a direct data source for Pharos, and thereby contributes to elevating targets to higher development levels, it is hard to estimate from the current coverage of understudied targets how many of what are now Tclin or Tchem targets received patent annotations while still in the understudied categories.

Patent bioactivities were found for six Tdark targets (Table 26). While none of them included active compounds, four of the targets were annotated with patent-exclusive compounds, three of them

with high scaffold and compound novelty. It remains to be seen whether these targets will continue to accumulate data and at some point will be elevated to another TDL.

Table 26: Tdark targets with patent annotations in ChEMBL 31.

UniProt accession	Name	IDG Family
Q6ZNB5	Putative short transient receptor potential channel 2-like protein	Other
Q8NHB7	Olfactory receptor 5K1	oGPCR
Q58A45	PAN2-PAN3 deadenylation complex subunit PAN3	Kinase
Q8N752	Casein kinase I isoform alpha-like	Kinase
Q92696	Geranylgeranyl transferase type-2 subunit alpha	Enzyme
Q96S38	Ribosomal protein S6 kinase delta-1	Kinase

157 Tbio targets had received patent annotations. For fourteen of them, that also included actives; one of these (the transcription factor Protein max) did not have any actives from other sources. For the other thirteen, the patent-exclusive actives did not show high compound novelty when compared to those annotated from other sources, but for ten targets their scaffold novelty was high. Looking at the general novelty of patent-exclusive compounds, regardless of activity, 38 of the targets were found in the highest novelty quadrant, compared to 7 with the lowest comparative novelty.

Tchem targets are commonly already considered well-studied, although there is certainly a degree of variation in how well they are truly explored. By definition, Tchem targets have at least one active compound. With 883 targets, they make up the largest part of the 1,370 human targets with patent annotations. In ChEMBL 31 (which does not include all actives that are included in Pharos, as that uses other data sources as well), 97 Tchem targets had actives only from patents, and a further 407 targets had actives from both patents and non-patent sources. Of the latter, the patent-exclusive actives were highly novel in respect to both scaffolds and compounds for 90 targets; only for 48 was there little gain in novelty from patent actives. When examining all compounds regardless of activity, high scaffold and compound novelty was found for 166 Tchem targets. Taken together, this shows that even for targets that aren't understudied, valuable data can be gleaned from patents.

Of the 379 Tchem targets where annotations from patents did not include actives, 347 have active compounds from non-patent literature; this leaves 32 Tchem targets where all actives that were reported elsewhere are absent from ChEMBL.

Most of the 320 Tclin targets with patent bioactivities had actives annotated from both patents and other sources. However, patent-exclusive actives were still highly novel for 43 of them, and novel in respect to either compounds or scaffolds for 16 and 156 Tclin targets, respectively. Across all patent-exclusive compounds, 38 targets had high novelty in both dimensions. Thus, even though this type of target may be less interesting for drug discovery and development than those for which

no approved treatment currently exists, there is still valuable information being added from patent bioactivity annotations.

Across the whole dataset, 246 targets had high scaffold and compound novelty when comparing all patent-exclusive compounds to those published elsewhere. The diversity of compounds annotated from patents, however, was not an indicator of their novelty: while 89 of the high novelty targets also had high diversity, 72 showed both low compound and scaffold diversity. A similar picture is seen for those 120 targets with low overall novelty, which also include some with high diversity (14 targets) or high scaffold and low compound diversity or the reverse (25 and 38 targets) besides those with low diversity (43 targets).

Focusing only on the 421 targets whose patent compounds displayed high compound and scaffold diversity, 19 did not have any compounds that did not also occur in non-patent literature, and thus have no novelty value (although they may have been new to ChEMBL at the time of their patent annotation and published in non-patent literature later, as the temporal aspect is not examined in this project). For 21 targets, the only source of compounds were patents, so their novelty can be regarded as high (Figure 38). This further reinforces the notion that high diversity does not have a strong influence on novelty.

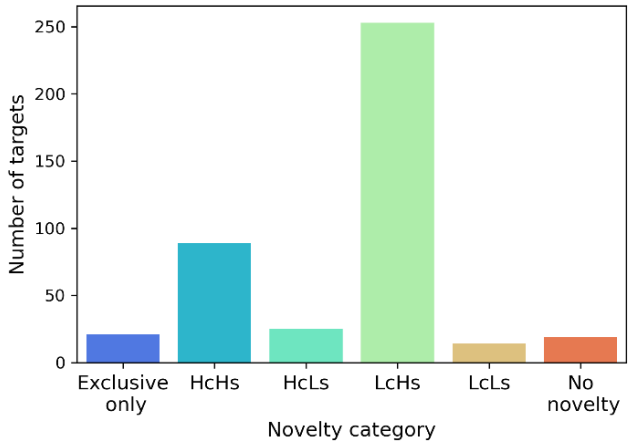


Figure 38: Distribution of targets among novelty categories, including targets with no compounds from non-patent sources and those with no novel compounds from patents. HcHs, high compound and scaffold novelty; HcLs, high compound, but low scaffold novelty; LcHs, low compound, but high scaffold novelty; LcLs, low compound and scaffold novelty.

Most patents from which bioactivities were annotated contained at least one novel compound. That was not the case for 197 targets: 20 of them had no patent-exclusive compounds from any patent, and 177 targets had at least one patent which did not contain exclusive compounds. Across the whole dataset, the number of patents annotated for each target had little bearing on its novelty or its diversity, as seen in Figure 39.

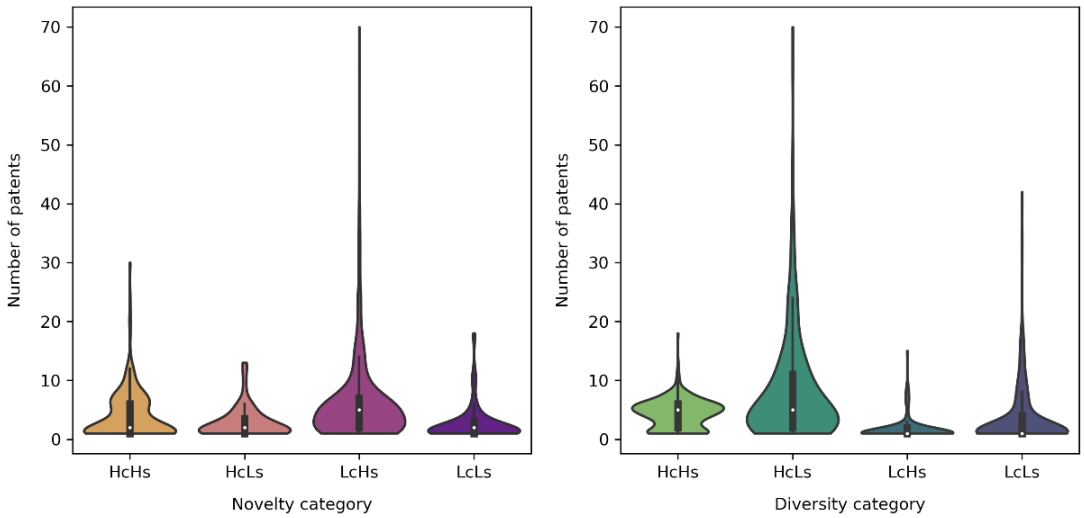


Figure 39: Number of patents per target from each novelty (left panel) and diversity category (right panel). HcHs, high compound and scaffold novelty/diversity; HcLs, high compound, but low scaffold novelty/diversity; LcHs, low compound, but high scaffold novelty/diversity; LcLs, low compound and scaffold novelty/diversity.

Apart from prioritizing understudied targets in general for patent bioactivity annotation, it may be worthwhile to try and identify where annotation efforts have yielded only low diversity or novelty.

140 targets did either not have any novel compounds from patents (20 targets) or were in the LcLs quadrant for novelty (Table 27). Most of them were Tchem or Tclin (81 and 49 targets), but also ten Tbio targets were among them. The most numerous target family with low novelty were enzymes (60 targets), followed by "other" targets and GPCRs. Kinases only had 5 targets with low novelty.

Table 27: Targets in the LcLs novelty quadrant, by family and TDL.

IDG Family	Number of targets	Tclin	Tchem	Tbio
Enzyme	60	17	40	3
GPCR	18	11	7	
Ion Channel	15	10	4	1
Transporter	7	2	5	
Nuclear receptor	6	6		
Kinase	5	1	3	1
Transcription factor	5		3	2
Epigenetic	5		5	
Other	19	2	14	3

While the novelty calculation relies on comparing compounds from both patent and non-patent sources and therefore can only be undertaken for targets where data from both sources exists, the diversity calculation can be applied to all of the 1,368 targets with sufficient data. 454 of them (33%) show both low compound and scaffold diversity, comprising 3 Tdark, 60 Tbio and 278 Tchem as well as 113 Tclin targets (Table 28). Again, enzymes make up the largest group, but here also kinases and "other" targets as well as GPCRs are numerous. The difference lies in the TDL,

where kinases and GPCRs have few targets in Tbio and Tdark, whereas these levels make up a substantial portion for enzymes and "other" targets.

Table 28: Targets in the LcLs diversity quadrant, by family and TDL.

IDG Family	Number of targets	Tclin	Tchem	Tbio	Tdark
Enzyme	135	12	103	23	1
Kinase	96	27	65	4	
GPCR	72	31	41		
Ion Channel	30	18	11	1	
Transporter	14	5	7	2	
Epigenetic	12	1	10	1	
Transcription factor	9		3	6	
Nuclear receptor	5	4	1		
oGPCR	1				1
TF-Epigenetic	1		1		
Other	79	16	39	23	1

Revisiting the coverage of understudied targets from Tdark and Tbio, first discussed in section 3.1, it is now possible to evaluate the data ChEMBL holds in more detail. Shown in Table 29 are the numbers for Tdark targets. Here, patents have not made a great contribution so far, with only 6 Tdark targets in total where bioactivities have been annotated from that source (see Table 26). However, three of the targets were categorized as high in novelty, and one had high diversity from patents. It is very likely that with increased efforts put into patent annotations, there are going to be changes seen in this category. However, when targets are better explored, they will by definition at some point be upgraded to Tbio or Tchem, and the current slim representation of Tdark targets from patent annotations is also a reflection of annotations that have helped to upgrade other Tdark targets.

Table 29: Data on Tdark targets in ChEMBL and the contribution of patents.

IDG Family	Pharos (Tdark)	Number in ChEMBL	Number from patents	High diversity (patents only)	High novelty
TF-Epigenetic	1	1			
Epigenetic	13	0			
Ion Channel	19	4			
Kinase	19	7	3	1	2
GPCR	30	17			
Transporter	79	1			
oGPCR	409	1	1		
Transcription Factor	414	2			
Enzyme	510	13	1		1
Other	4,185	26	1		

Tbio targets are more numerous than Tdark targets among patents but make up an even smaller percentage of all targets classified on that level in Pharos. The highest number of Tbio targets in patents per family is for the "other" category, with 54 targets (0.7% of all "other" targets), followed by 43 enzymes (1.6%) and kinases, where the 34 targets make up 21.1% (Table 30). Once again, kinases are in a special position. 12 of the 34 Tbio kinases with patent data also show a high diversity among their patent compounds, and 16 have high novelty values.

Table 30: Data on Tbio targets in ChEMBL and the contribution of patents.

IDG Family	Pharos (Tbio)	Number in ChEMBL	Number from patents	High diversity (patents only)	High novelty
Nuclear Receptor	7	7			
oGPCR	11	0			
TF-Epigenetic	29	7			
GPCR	103	74	6		3
Ion Channel	108	34	3	1	1
Epigenetic	133	49	2		1
Kinase	161	89	34	12	16
Transporter	278	47	3		
Transcription Factor	940	73	12	1	
Enzyme	2,725	569	43	13	11
Other	7,563	793	54	11	6

4 Conclusions

The aim of this project was to carry out the first systematic analysis of patent bioactivity data in ChEMBL. While there are many possible angles from which to approach this topic, the scope was limited to two points: the diversity of patent annotations, which allows estimating the breadth of data covered by patents so far, and their compound novelty, the amount of chemical space that has been opened up by patent data.

These analyses were approached using three diversity descriptor values, whose calculation was inspired by the work of González-Medina et al[32] for visualizing multi-dimensional chemical diversity. However, many modifications had to be made as the nature of the dataset analyzed here differed strongly from that in the original. Without doubt, further improvements could be made to the workflow described herein, if it were not for the limited scope of a thesis project, and with limited computational resources.

For example, another molecular framework definition than Bemis-Murcko scaffolds may be more suitable for the task. Especially when it comes to novelty calculations, the similarity of patent-exclusive and non-exclusive scaffolds would have been interesting to examine in more detail; here, making comparisons based on different levels of a scaffold tree[31,36] may have led to more insight.

Regarding compound diversity/novelty, there are many different fingerprints and distance calculations, and it may have been worthwhile to compare a few of them. Another point that is valid for all three descriptors but probably has the highest impact in compound diversity/novelty is the fact that only a mean value of all pairwise comparisons was used to describe the diversity/novelty of a target. Finding a way to capture the variability of the pairwise comparisons as well would surely lend more rigor to the analyses and make it easier to estimate the validity of data, for example for targets for which rather few compounds were annotated.

The physicochemical diversity and novelty value as calculated here turned out to be much less useful than anticipated due to several problems mentioned already throughout the results section. Some of the concerns could be remedied – for example, by exerting greater efforts to calculate or otherwise find descriptor values missing from ChEMBL, and thereby increasing the number of compounds used for calculation or applying a normalization beforehand. The selection of different physicochemical properties may also be an approach worth examining.

Despite these points, interesting insights were gained, as described in the summary chapter, and this project may prove useful as a starting point for developing a more fine-tuned analysis.

But which targets, then, should be prioritized for patent annotations? In current practice that is understudied targets, especially from the IDG priority list. Judging by development level and target

family, oGPCRs, transcription factors and “other” targets have the greatest fraction of Tdark/Tbio targets among them, followed by enzymes and transporters (Figure 4). Numerically, the “other” category has by far the highest number of Tdark and Tbio targets (4,185 and 7,563 targets), followed by enzymes and transcription factors (510/2,725 targets and 414/940 targets, respectively).

But target families are not considered of equal importance when it comes to clinical relevance, with a large fraction of approved drugs and compounds in development targeting only a handful of families. Thus, it makes sense to focus annotation efforts on the more “promising” target families, which is a notion underlying the selection of targets for the IDG priority list as well. This, however, may overlook important targets that do not belong to the traditionally most targeted classes. New or emerging insights that would lend more interest to previously neglected target families are also not considered with that approach. Patents applications are submitted for compounds modulating all sorts of targets, so research is clearly not restricted to kinases, ion channels and GPCRs, and it stands to reason that annotation efforts should not be, either.

5 References

1. Hughes, J. P., Rees, S. S., Kalindjian, S. B. & Philpott, K. L. Principles of early drug discovery. *Br. J. Pharmacol.* **162**, 1239–1249 (2011).

2. Lindsay, M. A. Target discovery. *Nat. Rev. Drug Discov.* **2**, 831–838 (2003).

3. Knowles, J. & Gromo, G. Target selection in drug discovery. *Nat. Rev. Drug Discov.* **2**, 63–69 (2003).

4. Agarwal, P. & Searls, D. B. Can literature analysis identify innovation drivers in drug discovery? *Nat. Rev. Drug Discov.* **8**, 865–878 (2009).

5. Oprea, T. I. *et al.* Unexplored therapeutic opportunities in the human genome. *Nat. Rev. Drug Discov.* **17**, 317–332 (2018).

6. Edwards, A. M. *et al.* Too Many Roads Not Taken. *Nature* **470**, 163–5 (2011).

7. Agarwal, P., Sanseau, P. & Cardon, L. R. Novelty in the target landscape of the pharmaceutical industry. *Nat. Rev. Drug Discov.* **12**, 575–576 (2013).

8. Santos, R. *et al.* A comprehensive map of molecular drug targets. *Nat. Rev. Drug Discov.* **16**, 19–34 (2016).

9. Hopkins, A. L. & Groom, C. R. The druggable genome. *Nat. Rev. Drug Discov.* **1**, 727–730 (2002).

10. Makley, L. N. & Gestwicki, J. E. Expanding the Number of 'Druggable' Targets: Non-Enzymes and Protein-Protein Interactions. *Chem. Biol. Drug Des.* **81**, 22–32 (2013).

11. Sheils, T. K. *et al.* TCRD and Pharos 2021: Mining the human proteome for disease biology. *Nucleic Acids Res.* **49**, D1334–D1346 (2021).

12. Online Mendelian Inheritance in Man, OMIM®. https://omim.org/.

13. Hajduk, P. J., Huth, J. R. & Tse, C. Predicting protein druggability. *Drug Discov. Today* **10**, 1675–1682 (2005).

14. Nguyen, D. T. *et al.* Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res.* **45**, D995–D1002 (2017).

15. Garbaccio, R. M. & Parmee, E. R. The Impact of Chemical Probes in Drug Discovery: A Pharmaceutical Industry Perspective. *Cell Chem. Biol.* **23**, 10–17 (2016).

16. Rodriguez-Esteban, R. & Bundschus, M. Text mining patents for biomedical knowledge. *Drug Discov. Today* **21**, 997–1002 (2016).

17. Magariños, M. P. *et al.* Illuminating the Druggable Genome through Patent Bioactivity Data. *PeerJ* (2023) doi:10.7717/peerj.15153.

18. Senger, S. Assessment of the significance of patent-derived information for the early identification of compound-target interaction hypotheses. *J. Cheminform.* **9**, 1–8 (2017).

19. Gadiya, Y., Zaliani, A., Gribbon, P. & Hofmann-Apitius, M. PEMT: a patent enrichment tool for drug discovery. *Bioinformatics* **39**, 1–3 (2023).

20. Tyrchan, C., Boström, J., Giordanetto, F., Winter, J. & Muresan, S. Exploiting structural information in patent specifications for key compound prediction. *J. Chem. Inf. Model.* **52**, 1480–1489 (2012).

21. Hattori, K., Wakabayashi, H. & Tamaki, K. Predicting key example compounds in

competitors' patent applications using structural information alone. *J. Chem. Inf. Model.* **48**, 135–142 (2008).

22. Falaguera, M. J. & Mestres, J. Identification of the Core Chemical Structure in SureChEMBL Patents. *J. Chem. Inf. Model.* **61**, 2241–2247 (2021).

23. Senger, S., Bartek, L., Papadatos, G. & Gaulton, A. Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents. *J. Cheminform.* **7**, 1–12 (2015).

24. Papadatos, G. *et al.* SureChEMBL: A large-scale, chemically annotated patent document database. *Nucleic Acids Res.* **44**, D1220–D1228 (2016).

25. Gaulton, A. *et al.* ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **40**, 1100–1107 (2012).

26. Mendez, D. *et al.* ChEMBL: Towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).

27. Papadatos, G., Gaulton, A., Hersey, A. & Overington, J. P. Activity, assay and target data curation and quality in the ChEMBL database. *J. Comput. Aided. Mol. Des.* **29**, 885–896 (2015).

28. Bento, A. P. *et al.* The ChEMBL bioactivity database: An update. *Nucleic Acids Res.* **42**, 1083–1090 (2014).

29. Gilson, M. K. *et al.* BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **44**, D1045–D1053 (2016).

30. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2017).

31. Langdon, S. R., Brown, N. & Blagg, J. Scaffold Diversity of Exemplified Medicinal Chemistry Space. *J. Chem. Inf. Model.* **51**, 2174–2185 (2011).

32. González-Medina, M., Prieto-Martínez, F. D., Owen, J. R. & Medina-Franco, J. L. Consensus Diversity Plots: a global diversity analysis of chemical libraries. *J. Cheminform.* **8**, 1–11 (2016).

33. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminform.* **7**, 1–13 (2015).

34. Bender, A. *et al.* How similar are similarity searching methods? A principal component analysis of molecular descriptor space. *J. Chem. Inf. Model.* **49**, 108–119 (2009).

35. Xu, Y. J. & Johnson, M. Using molecular equivalence numbers to visually explore structural features that distinguish chemical libraries. *J. Chem. Inf. Comput. Sci.* **42**, 912–926 (2002).

36. Schuffenhauer, A. *et al.* The scaffold tree - Visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* **47**, 47–58 (2007).

37. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).

38. RDKit: Open-source cheminformatics. doi:10.5281/zenodo.5242603.

39. Willett, P., Barnard, J. M. & Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **38**, 983–996 (1998).

40. IDG Understudied Proteins, Mar 28, 2022. https://druggablegenome.net/IDGProteinList.

41. Lewiecki, E. M. Role of sclerostin in bone and cartilage and its potential as a therapeutic target in bone diseases. *Ther. Adv. Musculoskelet. Dis.* **6**, 48–57 (2014).

42. Harding, S. D. *et al.* The IUPHAR/BPS guide to PHARMACOLOGY in 2022: Curating pharmacology for COVID-19, malaria and antibacterials. *Nucleic Acids Res.* **50**, D1282–D1294 (2022).

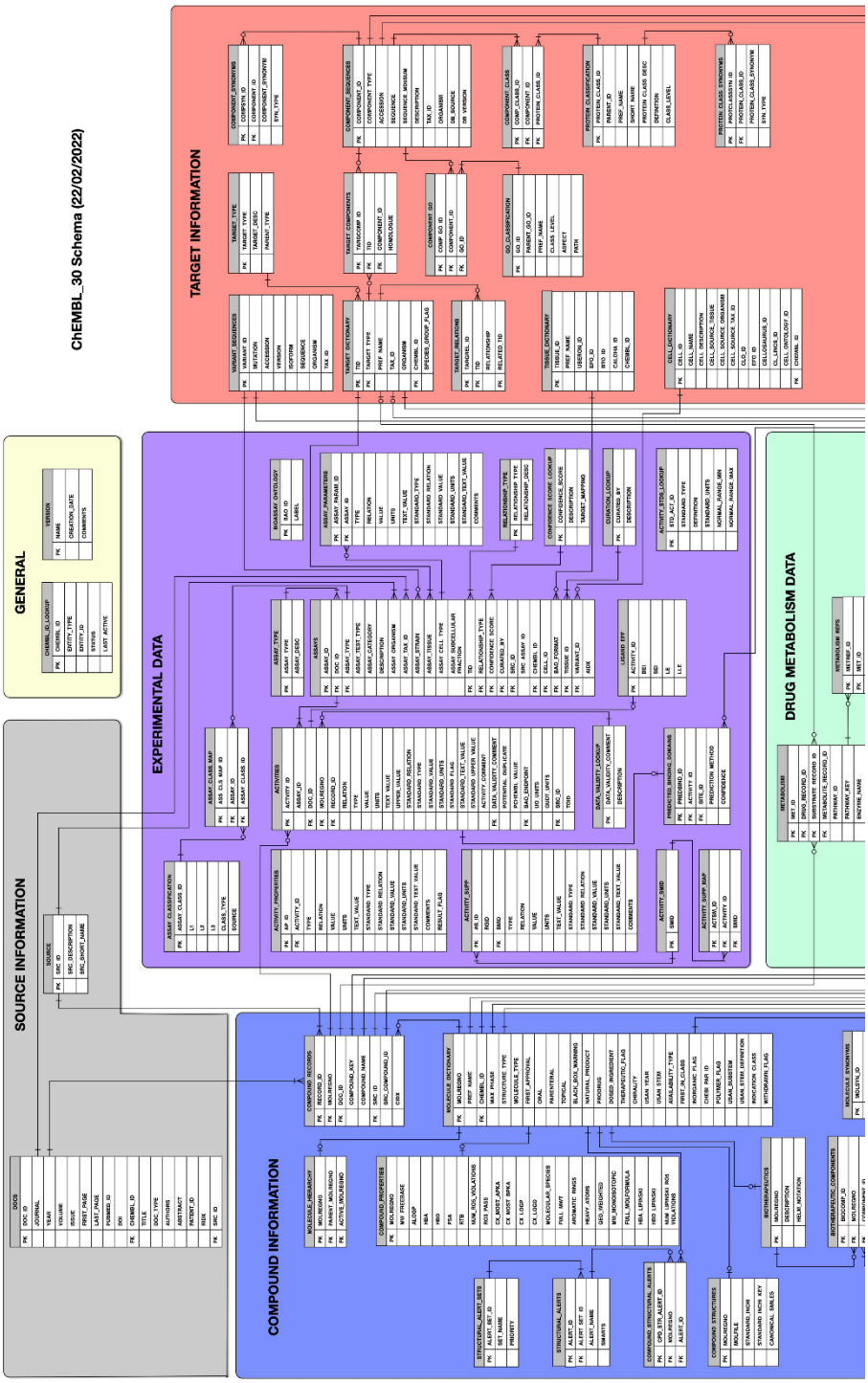
43. Kim, S. *et al.* PubChem 2023 update. *Nucleic Acids Res.* **51**, D1373–D1380 (2023).

44. Davis, M. I. *et al.* Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, (2011).

6 Appendix

6.1 ChEMBL database schema

The database schema in Figure 40 was valid for ChEMBL31 as no changes had been made to the structure since the previous release. An updated version for the current release is available on the ChEMBL website. The image was split in two here to improve legibility.



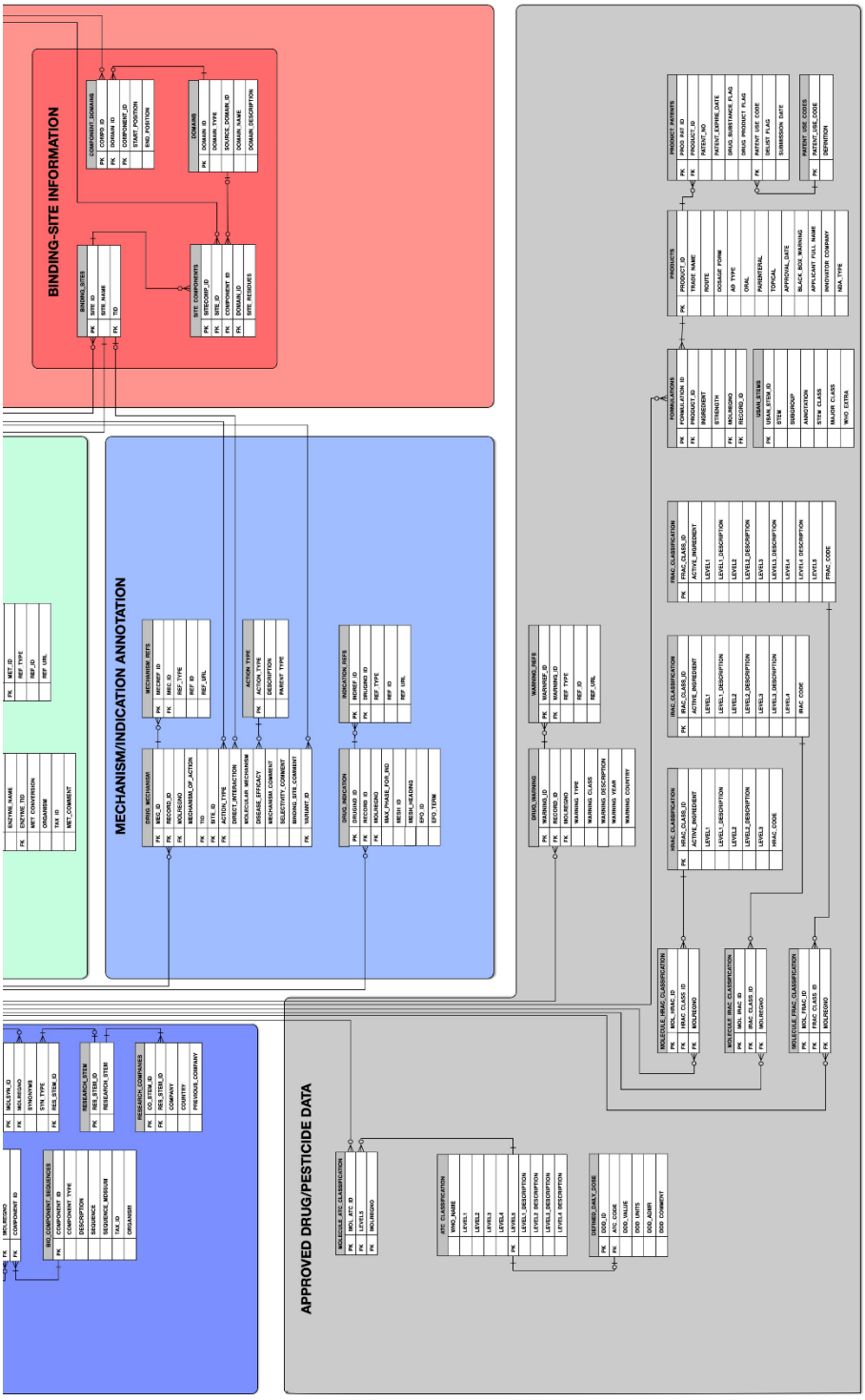


Figure 40: Database schema for ChEMBL30-31.