

**Illuminating the chemical space
of therapeutical relevance:
from pharmaceutical patents to
untargeted proteins**

Maria José Falaguera Mata

TESI DOCTORAL UPF / 2021

Thesis director:

Dr. Jordi Mestres

Departament de Ciències Experimentals i de la Salut



The research in this Thesis has been carried out at the Systems Pharmacology Group, within the Research Programme on Biomedical Informatics (GRIB) at the Barcelona Biomedical Research Park (PRBB).



RESEARCH
PROGRAMME
ON BIOMEDICAL
INFORMATICS



Barcelona
Biomedical
Research
Park

The research presented in this Thesis has been supported by a RETOS project from the Spanish Ministerio de Ciencia e Innovación (SAF2017-83614-R) and by Chemotargets S.L.



*«Mira al cielo
y cuenta las estrellas,
si puedes»*

*“Look up into the sky
and count the stars,
if you can”*

(Genesis 15:5)

Agradecimientos

A Jordi, por ser mi Virgilio en esta Divina Comedia a la catalana: ¡GRACIAS che!

A mi tía Reme y a mi tío Javier, por acogerme en vuestra casa como una hija más durante mis primeros años en Barcelona. Sin vuestra generosidad y vuestro cariño este sueño no habría sido posible.

A mi madrina, a mi tío Fali, a mis abuelos, a mi tita Concha, a mi tía Rosario, a Inma, a Rodrigo Padre y a Mapi: por vuestros valiosísimos consejos profesionales y personales y por creer siempre en mí.

A mis amigos del máster: Carmen, Gerard, las Marinas y Adri. Por ayudarme a llegar a ser la Doctora Falaguera *et al.*

A mis compañeros de laboratorio: Andreu, Xavi, Mariona, Juanma, Judith, David, Karolina, Audrey. En especial a Jordi, por aguantarme e incluso alimentarme en este tramo final, y a Joanna, por apoyarme y mirarme siempre con tanta admiración.

No me olvido de Alfons: gracias por tu constante disponibilidad durante estos años en lo profesional y también en lo personal.

A todo el equipo de Chemotargets, por enriquecerme con vuestros consejos, vuestra experiencia, vuestros conocimientos y vuestra profesionalidad. Sois unos *cracks*.

A mis nenas *barceloninas*: Clara, Berta, Gemma, Lore, Gise, Nùria, Anna, Glòria, Natàlia, Mirex. Por el apoyo, los consejos, las fiestas y las risas. Por hacer que me enamore perdidamente de esta ciudad. Esta tesis es también muy vuestra.

A mis *guadas*, por crecer juntas y alegrarse tanto de mis éxitos como si fueran los suyos propios.

A Rodrigo, por todo lo que hemos pasado juntos estos años y por la ilusión de lo que nos queda por vivir. Agárrate. ;)

A mis padres y a mi hermana, por hacerme llegar hasta aquí. Sin vosotros, de verdad, nada de esto habría sido posible. Por todo vuestro esfuerzo, vuestra entrega, vuestro apoyo, vuestro amor: Gracias.

Por último a Dios, que me acompaña siempre. Por tantas luces en los momentos malos. No dejes que pierda nunca la curiosidad y la capacidad de asombro. Esta tesis es para Ti.

Abstract

Systems pharmacology is the discipline that studies the so-called ‘pharmacological space’ with a holistic and network-based perspective. Its challenge is to shed light on the astonishingly sophisticated biological processes that characterize living cells, and the effect that exogenous chemical entities with therapeutic purposes have when entering them. From this perspective, a series of novel computationally-developed methodologies are presented in this Thesis. They aim from the exploration of the pharmacologically-relevant chemical space claimed in patents, to the unveiling of pharmacological opportunities to drug yet untargeted proteins.

Resumen

La Farmacología de Sistemas es la disciplina que estudia el llamado «espacio farmacológico» desde una perspectiva holística y basada en redes. Su reto consiste en arrojar luz a los deslumbrantemente sofisticados procesos biológicos que caracterizan las células vivas y el efecto que tienen entidades químicas exógenas con fines terapéuticos al entrar en ellas. Desde esta perspectiva se presentan en esta tesis una serie de novedosas metodologías desarrolladas computacionalmente. Éstas pretenden desde la exploración del espacio químico farmacológicamente relevante reclamado en patentes, hasta el descubrimiento de oportunidades farmacológicas para atacar proteínas para las cuales aún no se ha encontrado un fármaco.

Preface

In the era of *Big Data* with zettabytes of data/information within our reach, devising ways to structure, organise, integrate and ultimately query them is an utmost need to unveil the cosmos of knowledge they enclose. Data generated in the context of Biology defines what is known as *biological space* and its study from a holistic and network-based perspective has given rise to the new discipline of *Systems Biology*. Similarly, data generated in the context of Chemistry defines the *chemical space* which is the object of study of *Systems Chemistry*. In the area where these two spaces converge, which could be termed as the *pharmacological space*, the novel discipline of *Systems Pharmacology* has emerged. Its challenge is to shed light on the astonishingly sophisticated biological processes that characterize living cells, and the effect that exogenous chemical entities with therapeutic purposes have when entering them.

With this vision, the main objective of the present Thesis was the development of new methods and tools that contribute to explore pharmacological opportunities at both the chemical and the biological space. The document has been divided into six parts. The first part provides an overview of current state of the art of the system-based disciplines mentioned. The next one introduces the primary objectives pursued. The third part compiles the two publications and the two manuscripts in preparation that have resulted from this Thesis. Finally, the last three parts discuss the results obtained, list the main conclusions derived, and provide a general list of relevant references, respectively.

Table of contents

	Pag.
Abstract	ix
Preface	xi
List of publications	xv
Part I: Introduction	1
I.1 The chemical space	3
I.2 Biologically-active chemical space	6
I.3 Pharmacologically-active chemical space	9
I.4 Chemical series and privileged structures	11
I.5 Patent-derived chemical data and SureChEMBL database	14
I.6 Chemoinformatics tools	19
I.7 The biological space	25
I.8 The proteome space	30
I.9 The Illuminating the Druggable Genome initiative	32
I.10 Polypharmacology in drug discovery	35
Part II: Objectives	39
Part III: Results	43
III.1 Identification of the core chemical structure in SureChEMBL patents	45
III.2 Congenericity of claimed compounds in patent applications	87

III.3 Illuminating the chemical space of untargeted proteins	117
III.4 A map of the proteome targetable by dual-acting agents	145
Part IV: Discussion	167
IV.1 Patents as a source of novel chemical space	169
IV.2 The chemical scope of patent applications	170
IV.3 Illuminating the chemical space of untargeted proteins	171
IV.4 A novel ontology to help DADs discovery	172
Part V: Conclusions	173
Part VI: References	177
Appendix	189

List of publications

Articles:

- Falaguera, M. J. & Mestres, J. Identification of the core chemical structure in SureChEMBL patents. *Journal of Chemical Information and Modeling* **2021**, *61*, 2241–2247.
Journal Impact Factor: 4.956.
- Falaguera, M. J. & Mestres, J. Congenericity of claimed compounds in patent applications. *Molecules* **2021**, *26*, 5253. *Special Issue dedicated to Prof. Jürgen Bajorath*.
Journal Impact Factor: 4.411; Citations: 1.
- Falaguera, M. J. & Mestres, J. Illuminating the chemical space of untargeted proteins. To be submitted.
- Falaguera, M. J. & Mestres, J. A map of the proteome targetable by dual-acting agents. In preparation.

Poster communications:

- Falaguera, M. J. & Mestres, J. Identification of the core chemical structure in SureChEMBL patents. Poster communication presented at the Symposium to Celebrate 10 Years of the ChEMBL Database. 2019. Hinxton (UK).

Part I: Introduction

I.1 The chemical space

Analogous to the cosmic space, with around 200 billion trillion stars grouped in around 2 trillion galaxies¹ and new nascent ones in continuous forming process (Figure 1), the concept of **chemical space** refers to the ensemble of all possible molecular entities, both naturally occurring and artificially synthesized in a laboratory, which should be considered when searching for a new drug². Current theoretical estimations of the chemical space size point at orders of magnitude of 10^{60} organic molecules showing the chemical and physical properties necessary to be likely orally active drugs.^{2,3}



Figure 1. *The Pillars of Creation.* Photography taken by the Hubble Space Telescope of towers of cosmic gas and dust in a star-forming region of the Eagle Nebula. Figure extracted from NASA.⁴

Introduction

In the early era of pharmacology, back in the mid-19th century, the known chemical space was mainly constituted by active ingredients extracted from medicinal plants to be tested *in vivo* in model animals, such as morphine, isolated from opium extract, and papaverin with antispasmodic properties.⁵ More recently, with the molecular biology revolution and the explosion of **combinatorial chemistry** and **high-throughput screening** (HTS) programmes,⁶ the amount of synthetic compounds produced in mass to be tested *in vitro* against isolated macromolecular targets increased dramatically the known chemical space size with almost 400 chemical libraries of 10,000 to 100,000 compounds each one⁷ produced every year.⁸

In an attempt to organize this chemical data and make it accessible for the scientific community, the first **chemical open-access databases**, ChEBI (Chemical Entities of Biological Interest)⁹ and PubChem,¹⁰ were launched in 2004 by the European Molecular Biology Laboratory (EMBL) and the National Center for Biotechnology Information (NCBI), respectively. This was followed by the release of several other public chemical repositories built with a medicinal chemistry focus that have been updated with new upcoming chemical information since then (Table 1). Together with the structure and molecular information of chemical compounds, most of these repositories also contain data related to the interactions these molecules are known to have over specific biological systems. This information is used by drug discoverers as a map to improve their understanding of chemical **structure-activity relationships (SAR)** and to identify better pharmacological opportunities.

Table 1. Public databases of the known chemical space.

Name	Size	Description (Website)
ChEBI ⁹	0.03 M	Dictionary of molecular entities focused on small chemical compounds. (https://www.ebi.ac.uk/chebi/)
PubChem ¹⁰	110.0 M	Known molecules from various public sources. (http://pubchem.ncbi.nlm.nih.gov)
ChemSpider ¹¹	100.0 M	Online resource from the Royal Society of Chemistry. (http://www.chemspider.com/)
ZINC ¹²	750.0 M	Commercially-available small molecules. (http://zinc.docking.org)
BindingDB ¹³	1.0 M	Bioactive molecules with binding affinity data. (http://www.bindingdb.org)
ChEMBL ¹⁴	2.1 M	Small molecules annotated with experimental bioactivity data. (https://www.ebi.ac.uk/chembl/db)
DrugBank ¹⁵	0.5 M	Experimental and approved small molecule drugs. (http://www.drugbank.ca)
DrugCentral ¹⁶	0.001 M	Experimental and approved small molecule drugs. (https://drugcentral.org/)
GtoPdb ¹⁷		Expert-curated database of molecular interactions between ligands and their targets. (https://www.guidetopharmacology.org)
SureChEMBL ¹⁸	17.0 M	Small molecules extracted from patents by text- and image-mining techniques. (https://www.surechembl.org/search)

I.2 Biologically-active chemical space

Despite the vastness of the chemical space, the interest of drug discovery (DD) researchers is mainly focused on the regions that are occupied by biologically-active compounds, which are the ones showing potentially-therapeutic characteristics.^{7,19} These characteristics include topological and physicochemical properties that allow for the specific binding interaction with the molecular recognition patterns on biological molecules implicated in disease phenotypes, such as proteins, RNAs and DNAs.¹⁹

To assess the distribution of this bioactive chemical space throughout the whole chemical space, multidimensional representations in form of chemographic maps have been proposed.²⁰ These are global positioning systems, built on the basis of SAR, that involve mapping compounds into coordinates of chemical descriptors derived from their topological and physicochemical properties, in order to group together molecules with similar structures, and thus similar bioactivities. These representations solve the dimensionality problem of comparing and visualizing multiple molecular properties at once in two- or three-dimensional maps by reducing them to a series of molecular descriptors using algorithms such as the self-organized maps (SOM) or the principal component analysis (PCA, Figure 2).²¹

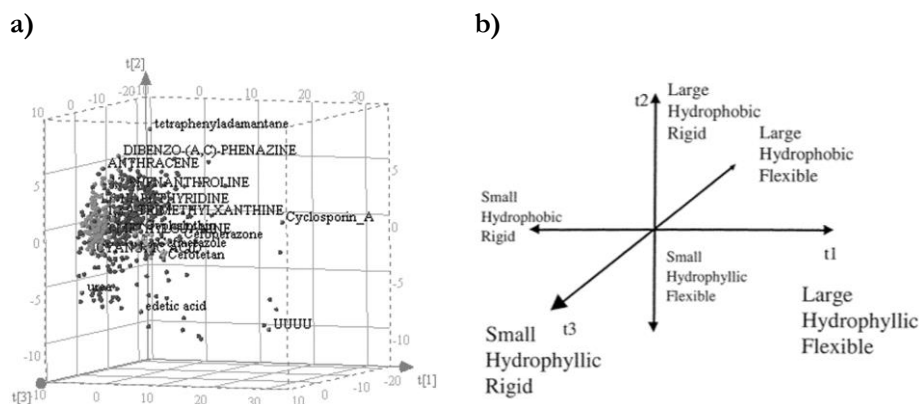
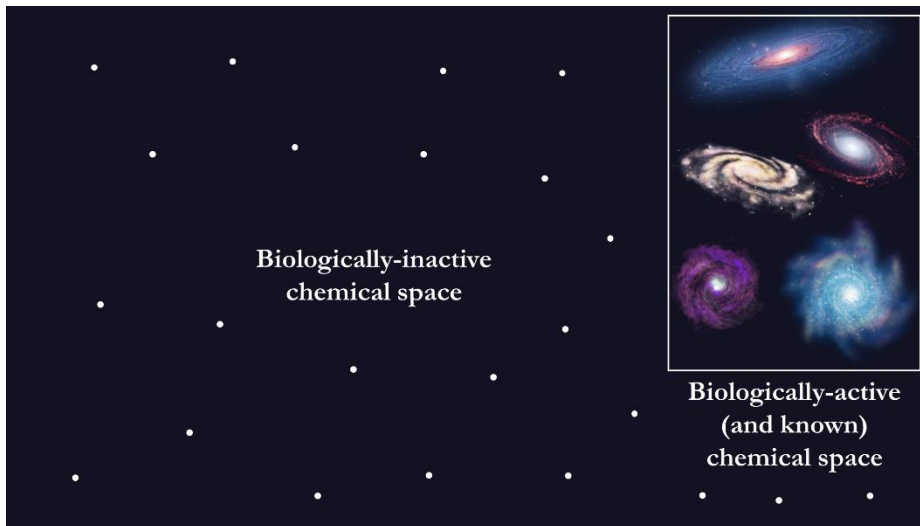


Figure 2. a) Projection of a dataset of molecules on a three-dimensional PCA plot. b) Translation of the three principal components in (a) into interpretable chemical descriptors. Figure adapted from Oprea and Gottfries (2001).²⁰

Current chemographic maps built with the data generated during a century of medicinal chemistry and HTS programmes show that the known biologically-relevant compounds are not sparse throughout all the chemical universe but clustered together in small and discrete regions of it enriched with compounds of similar structure that bind to similar targets, being reminiscent to galaxies. The question that remains, is whether these galaxies of known bioactive molecules are the only ones in the chemical space worthy to be explored for therapeutic opportunities, or whether they are only a subpart of it and we are way far from fully unveiling the likely bioactive chemical space (Figure 3).¹⁹

Introduction

a)



b)



Figure 3. Possible distributions of the biologically-active chemical space. **a)** Full biologically-active chemical space is already known (galaxies). **b)** Only a subpart of the biologically-active chemical space is known (galaxies), with other regions yet to be discovered (white stars).

I.3 Pharmacologically-active chemical space

Continuing with the cosmic analogy, biologically-active compounds can also be seen as planets that possess the basic suitable conditions to harbour life. But, like having liquid water does not ensure the planet habitability, the capacity of a compound to potently bind a biological target *in vitro* is not enough to be a drug but it constitutes only the first step in a typical DD process. **Hit** compounds, as this biologically-active compounds are known, need to be then tested for its potential **bioavailability**, this is, its potential susceptibility to be processed as a drug by the human body according to pharmacokinetics. Decades of successes and failures in the approval of promising hits as new drugs, have led to some basic rules to describe the molecular properties important for orally-administered drugs bioavailability. The most important ones have been encapsulated in the **Lipinski's rule of five (RO5)**, so named in honour of the first one to define them in 1997,²² also known as the **'ADME' properties**: absorption, distribution, metabolism and excretion. The **absorption** property refers to the orally-administered compound capacity of being aqueous soluble enough to permeate the mucosa surfaces in the digestive tract and be taken into the bloodstream. Once in the bloodstream the compound needs to be carried to its effector site by **distribution** and transfer from one body compartment to another and this depends on its molecular size, polarity and binding to serum proteins capacity. Distribution can be a serious problem at some natural barriers like the blood-brain barrier. The third ADME property is the drug **metabolism**. Compounds begin to break down as soon as they enter the body but the majority of small-molecule drugs metabolism is carried out in the liver by cytochrome P450 enzymes. They attack the vulnerable chemical functionalities of the initial compound to convert it into

Introduction

new compounds called metabolites which should be pharmacologically-active for a useful *in vivo* effect.¹⁹ Finally, initial compounds and their metabolites need to be removed from the body via **excretion** to avoid the accumulation of foreign substances that may adversely affect normal metabolism. Hydrophobic drugs, to be excreted, must undergo metabolic modifications making them more polar, while hydrophilic drugs can undergo excretion directly, without the need for metabolic changes to their molecular structures. The main excretion via are the kidneys by the urine, and others include the gut by the faeces and the lungs by anaesthetic gases.

Hit compounds optimized in their ADME properties are named **'leads'** and the limited range of leads' molecular properties define the areas of the chemical space where orally administered drugs are more likely to reside, known as the **pharmacologically-active chemical space**. This notwithstanding, despite candidate drugs that conform to the RO5 tend to have lower attrition rates during clinical trials,²³ the fact that a molecule fulfils these rules does not always guarantee its drug-likeness¹⁹ and it must be borne in mind that it can always be place for failure in the following DD phases, which include the pre-clinical testing in animal models and the clinical development in humans to assess both drug efficacy and safety *in vivo* (Figure 4).

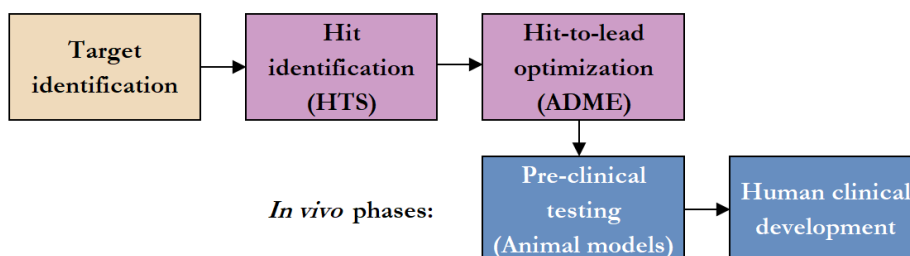


Figure 4. Scheme of the drug discovery and development process.

I.4 Chemical series and privileged structures

Biologically-active and pharmacologically-active chemical spaces can be organized in **chemical series**, defining collections of structurally-analogue compounds with similar biological annotations. This chemical series concept is very widely used in DD projects today since the seed step in most of them is precisely the identification of chemical series of hit compounds showing promising affinities for therapeutic targets. Apart from with chemographic coordinates, as shown in Chapter I.2, these chemical sets can also be characterized using more human-intuitive approaches that put the focus on structural properties. These approaches consist of the extraction of the most repeated substructures found in the compounds, named the '**privileged structures**', as the most likely ones to contain the structural properties responsible for the molecules bioactivity against the target (Figure 5). The term was first introduced by Evans *et al.* (1988)²⁴ as 'single molecular frameworks able to provide ligands for diverse receptors', and it was exemplified with the benzodiazepines framework which is present in several types of central nervous system (CNS) agents.²⁵ Nowadays, the definition has evolved to a more generic concept that includes all the structures very frequently occurring in a given chemical series of bioactive molecules.

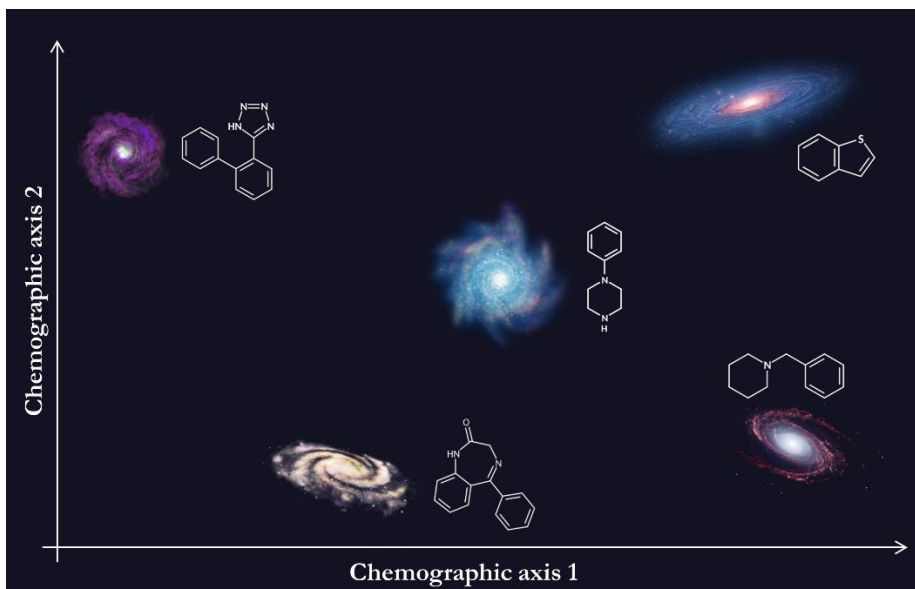


Figure 5. Chemical series of bioactive structurally-analogue compounds, here represented as galaxies, can be described using chemographic coordinates or privileged structures.

Like its definition, the way to obtain the privileged structures of a chemical series is quite versatile with many chemoinformatic groups worldwide proposing their own. Reviewing current published approaches, I concluded that they can be summarised in three categories (Figure 6): (i) approaches based on the extraction of the Murcko scaffolds²⁶ shared between the molecules, (ii) approaches based on the assessment of maximum common substructures (MCS) found between the molecules,²⁷ and (iii) approaches based on the identification of common fragment-based structures derived from the retrosynthetic fragmentation of the molecules (privileged fragments), which is the basis of some popular algorithms for chemical series characterization, like the matched molecular pairs (MMP) one.²⁸

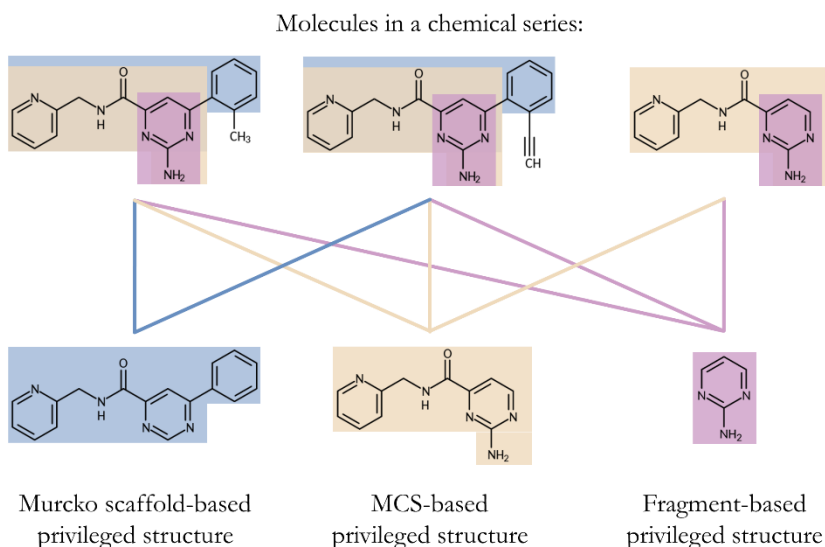


Figure 6. Examples of privileged structures calculation for a chemical series.

Once the privileged structures in a collection of hit compounds with bioactivity for a given target are identified, they are commonly used in the following DD steps to design **focused chemical libraries**, a.k.a. targeted chemical libraries,^{29,30} to be experimentally tested against the target in order to enhance hit discovery rates. These libraries are enriched with other compounds containing the privileged cores as well, but with variations in their side chains decoration with respect to the initial ones. The diversity coverage of chemical libraries built on the basis of privileged structures will depend on the size and complexity of the privileged structures chosen. Large and complex structures will result in panels of highly congeneric molecules, while small fragment-based ones will give rise to more diverse molecular collections. The election between a congeneric panel and a more diversity-covering one will ultimately depend on the purpose of the screening.

I.5 Patent-derived chemical data and SureChEMBL database

Chemical series of bioactive compounds for therapeutic targets are reported in public literature and in patents, being the later one the first line source of novel chemical data of biological relevance. As soon as a new lead compound is identified, discoverers claim for legal protection of their finding through a **patent application** document (there are some exceptions) containing a very detailed technical description³¹ that includes the chemical structure of the compound or series of compounds claimed, their synthesis process and some bioannotations for them. Also, as an abstract representation of the compounds claimed, applicants normally include a **Markush structure** (Figure 7).

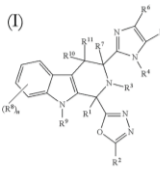
<p>Oxadiazole beta carboline derivatives as antidiabetic compounds</p> <p>Patent ID: US-8754099-B2</p> <p>Inventor: Liangqin Guo, William K. Hagmann, Shuwen He, Zhong Lai, Jian Liu, Ravi P.</p> <p>Current assignee: Merck Sharp and Dohme Corp</p> <p>Date of patent*: 2014-06-17</p> <p>Classification: C07D471/04 (Ortho-condensed systems) A61P3/00 (Drugs for disorders of the metabolism)</p>	<p style="text-align: center;">Abstract:</p> <div style="display: flex; align-items: center;"> <div style="flex: 1;"> <p>(I)</p>  <p style="text-align: center;"><i>Markush structure</i></p> </div> <div style="flex: 2; padding-left: 10px;"> <p>Beta-carboline derivatives of structural formula I are selective antagonists of the somatostatin subtype receptor 3 (SSTR3) and are useful for the treatment of Type 2 diabetes mellitus and [...] depression and anxiety.</p> </div> </div>
---	---

Figure 7. Example of a chemical patent application extract. *Date of patent application granted.

Compared to literature-derived chemical data, only 3% of the compounds claimed in patents are then reported in scientific publications. And for this 3%, the average *lag* time between patent deposition and scholarly literature report goes from 3 to 4 years.³² Early access to this patent-derived data is quite easy using commercial databases such as Excelra GOSTAR, Thomson Reuter Pharma and Elsevier Reaxys, which provide regularly-updated and high-quality chemical information extracted from patents using automatized text- and image-mining techniques and later manual-curation (Table 2). Alternatively, there are public counterparts offering open-access data, such as SCRIpDB or ChEBI (Table 2), but they have very limited patent and chemical coverages in comparison and have had few or none data updates since years (Table 2). However, this scenario changed drastically in 2016 with the release of the first open-access and weekly-updated database of chemical information extracted from patents by text- and image-mining techniques, **SureChEMBL database**.¹⁸ Derived from a commercial chemistry patent mining product originally developed as SureChem by Digital Science Ltd, SureChEMBL was acquired by the EMBL-EBI in 2013 with the remit to expose full functionality and underlying chemical structure content to the public domain.³³ Following a short migration period, the first version was released (April 2016) containing 17 million compounds extracted from 14 million patent documents (and their attached MOL files if available) from all four major patent authorities, namely, the European Patent Office (EPO, <https://www.epo.org>), the United States Patent and Trademark Office (USPTO, <http://www.uspto.gov>), the Japanese Patent Office (JPO, <https://www.jpo.go.jp>) and the World Intellectual Property Organization (WIPO, <http://www.wipo.int>) (Figure 8).

SureChEMBL database represents a huge step forward in the early and free access by the whole scientific community to novel chemical space of biological relevance. Given the time and scope penalties associated to literature-derived chemical data in comparison and the significance of patent-derived information for early identification of compound-target interaction hypotheses, in the future, more DD investments should be devoted to the development and provision of databases and tools to search patent information in a more comprehensive, reliable, and efficient manner.³⁴

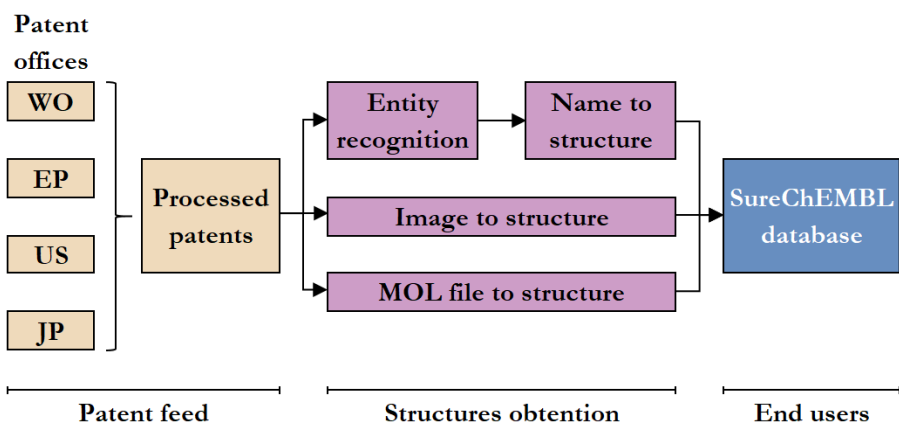


Figure 8. Overview of SureChEMBL data pipeline from the raw patent feed to the standardized compounds in the database. Figure adapted from Papadatos *et al.* (2016).¹⁸

Table 2. Databases of patent-derived chemical space.

Name	Description (Website)
<i>Commercial databases</i>	
PatentPak (CAS SciFinder)	Product of CAS SciFinder that provides access to chemical data extracted from 18 million patents. Daily-updated. (https://scifinder.cas.org)
GOSTAR (Excelra)	Largest manually-curated resource of SAR data extracted from 2,900 patents and 3,400 papers. (https://www.gostardb.com)
Reaxys (Elsevier)	It contains chemical data extracted from 2 million patents from 105 patent offices. (https://www.elsevier.com/solutions/reaxys)
Derwent (Clarivate)	Database derived from Thomson Reuter Pharma for searching and analysing chemical patents. (https://clarivate.com/derwent)

Table 2. (continued)

Name	Description (Website)
<i>Public databases</i>	
ChEBI ⁹	It contains chemistry automatically extracted from the title and abstract of a subset of biologically-relevant EPO patent documents. Data has not been updated in the last years. (https://www.ebi.ac.uk/chebi)
SRIPDB ³⁵	Database containing >10 million compounds automatically extracted from the attached MOL files of granted US patents published between 2001 and 2011. (http://dcv.uhnres.utoronto.ca/SCRIPDB)
IBM contribution to the NIH ^{*,36}	It contains >2 million chemical structures extracted from 4 million full-text patents (EPO, WIPO, USPTO). Data has not been updated since 2011 and is deposited in PubChem.
UniChem ³⁷	Repository that integrates several sources of patent chemistry, such as IBM, ChEBI, SCRIPDB and Thomson Pharma. (https://www.ebi.ac.uk/unichem)
SureChEMBL ¹⁸	Large-scale, chemically-annotated, up-to-date database that contains 17 million compounds extracted from 14 million patents (EPO, WIPO, USPTO, JPO) by text- and image-mining techniques. (https://www.surechembl.org/search)

*US National Institutes of Health

I.6 Chemoinformatics tools

In order to close the chemical space section, I will now describe the main tools employed nowadays in chemoinformatics to process and analyse chemical data. To computationally integrate and compare chemical data from different sources, chemical structures need to be identified and represented in a standardized and transferable manner. Today, different approaches have been taken to address this issue. Since 2005, the IUPAC (International Union of Pure and Applied Chemistry) proposes the **InChI** (IUPAC International Chemical Identifier) as a textual unique identifier for chemical substances.³⁸ To get the InChI of a compound, the process starts with the normalization of its structure, removing redundant information, and then it follows with the canonicalization into a unique form, such that any representation of this compound would collide into a single unique graph representation. This canonical representation is serialized into a textual form containing six different layers of information related with the structure, the charges, the stereo chemistry and other chemical features of the compound. When this InChI was found to be too long to be efficiently searched and stored, the InChI keys were developed. For their calculation, InChI string is hashed into a 25 characters length alphanumeric code were 14 of this characters result from the connectivity information of the InChI followed consecutively by a hyphen and 8 more characters resulting from the remaining layers of the InChI. After this, a single character indicating the version of InChI used and a single checksum character are found. The chance for two different compounds to have the same InChI key is estimated in 1.3 for every 10^9 compounds, meaning a single collision into 75 databases of 10^9 compounds each. A more human-readable representation of chemical structures than InChI are the simplified molecular-input line-

Introduction

entry system (**SMILES**), a specification in form of a line notation for describing the structure of chemical species using short ASCII strings. The original SMILES specification was initiated in the 1980s by David Weininger at the USEPA Mid-Continent Ecology Division Laboratory in Duluth.^{39,40,41} It has since been modified and extended by others, most notably by Daylight Chemical Information Systems.⁴² In contrast with InChI, there are usually a large number of valid generic SMILES which represent a given structure, thus, a canonicalization algorithm exists to generate one special generic SMILES among all valid possibilities; this special one is known as the 'unique SMILES'.⁴² Related to SMILES, the SMILES arbitrary target specification (**SMARTS**) is a line notation for specifying substructural patterns in molecules.⁴³ Another way to transfer molecular information is using chemical table files (CT files), such as the popular MOL file format (**molfile**) and the structure-data file (**SDF**) format developed by MDL Information Systems (MDL), which was acquired by Symyx Technologies, then merged with Accelrys Corp., and now called BIOVIA, a subsidiary of Dassault Systèmes of Dassault Group.⁴⁴ The molfile consists of some header information about the molecule, the 'Connection Table' (CT) listing each atom in the molecule in its x-y-z space coordinates, and then the information of the bonds connecting atoms, followed by other sections for more complex information.⁴⁵ SDF files wrap multiple molfiles of different compounds and join them using lines of four dollar signs (\$\$\$\$) as delimiters. A feature of the SDF format is its ability to include associated data of the molecules represented.

To determine the similarity between two molecules, their structures can be encoded in the named molecular fingerprints and then compared calculating their level of molecular similarity. The most common types of fingerprints consist of a series of binary digits (bits) that represent the

presence or absence of particular substructures in the molecule like a boolean array.^{46,47} Topological or **path-based fingerprints** index the small molecule fragments based on linear segments of up to 7 atoms, ignoring single atom fragments of 'C', 'N', and 'O'. A fragment is terminated when the atoms form a ring. For each of these fragments, the atoms, bonding and whether they constitute a complete ring is recorded and saved in a set so that there is only one of each fragment type. Chemically identical versions, i.e. ones with the atoms listed in reverse order and rings listed starting at different atoms, are identified and only a single canonical fragment is retained. Each remaining fragment is assigned a hash number from 0 to 1020 which is used to set a bit in a 1024 bit vector. On the other hand, **Atom pair fingerprints** are constructed by extracting the shortest path between all pairs of atoms in a small molecule, encoding the paths with descriptors of the atom types, the number of bonds for both atoms and their topological distance. The descriptors are then converted into bit strings, which are subsequently concatenated into one number. This number is hashed into the index space and its corresponding position in the fingerprint set is 1.⁴⁸ **Topological Torsion** fingerprints (TT) are calculated in essentially the same way but considering as fragments four consecutive bonded non-hydrogen atoms along with the number of non-hydrogen branches.⁴⁹ Finally, circular or **Extended Connectivity Fingerprints** (ECFPs), firstly introduced in 2000 by Dassault Systèmes, offer a number of advantages over the other schemes. Each atom in a molecule can be viewed as the centre of a radius of perception or 'orbit'. Information about an atom can be iteratively gathered by first examining immediate neighbours, then the neighbours of those neighbours, and so on. An ECFP is defined as the set of all atom identifiers for each radius of perception up to the limit n . As the radius of perception expands (n increases), this set includes all identifiers found in both previous

Introduction

iterations and the current one. Typical ECFPs used are ECFP4s corresponding to a radius of perception of 2 (diameter of perception 4). The features that make ECFPs especially useful compared to other options are that the calculation algorithm is composed of very few units, making implementation straightforward; that many variations on the base algorithm are possible conferring it flexibility to be optimized for different uses; that they are not predefined and can represent an essentially infinite number of different molecular features (including stereochemical information); and that their features represent the presence of particular substructures, allowing easier interpretation of analysis results.^{50,51} Once obtained the fingerprints of a pair of molecules, no matter which type, the similarity between them can be calculated using similarity metrics like the **Tanimoto (or Jaccard) coefficient**, defined as the ratio of the number of features common to both molecules relative to the total number of features,⁵² or the **Dice coefficient** (Hodgkins-Richards index), defined as the number of features in common to both molecules relative to the average size of the total number of features present.⁵³ They range from 0.0 to 1.0 inclusive.

Today, several software solutions are available to translate compound libraries from one format to another, compute molecular fingerprints and calculate similarities between them. Of mention are OpenBabel,⁵⁴ ChemAxon⁵⁵ and the very popular open-source **RDKit**⁵⁶ which can be easily imported and used as a Python library. In Figure 9 a summary of the main molecular identifiers and descriptors mentioned are exemplified.

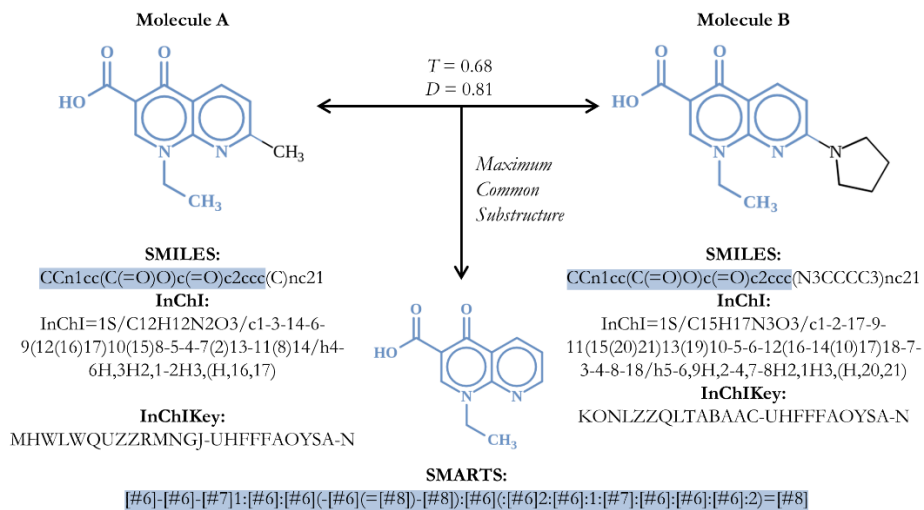


Figure 9. Example of molecular formats and descriptors. Maximum common substructure highlighted in blue. T , Tanimoto similarity between ECFP4 fingerprints. D , Dice similarity between ECFP4 fingerprints.

I.7 The biological space

The historical feat achieved in 2001 by the **Human Genome Project** (HGP) of sequencing and assembling the euchromatic portion of the human genome (Figure 10),^{57,58} could be comparable to an outward exploration of a planet, a galaxy or the cosmos, since it was an unprecedented inward voyage to discover the nature's complete genetic blueprint for building a human being (Figure 11). This international effort, carried out by the biotech company, Celera Genomics, and the International Human Genome Sequencing Consortium,⁵⁸ unveiled a huge and complex genetic cosmos until then unknown of size ~ 3 billion base pairs (bp) and ~ 30 thousand protein-coding human transcripts.⁵⁷ Provided since then to the whole scientific community, it serves as a guide to help understand the human genetic instructions book and to explode this knowledge for therapeutic purposes.



Figure 10. The Human Genome publication on Science⁵⁷ and Nature⁵⁸ covers.

Introduction

For the 20th anniversary of the publication of the first draft of the human genome that we are celebrating this year, special issues in very relevant scientific journals in the field, such as *Science* and *Nature*,⁵⁹ have been published reviewing the impact of this achievement in fields like genomics, drug discovery, medicine and scientific literature since 2001. Analysis of the evolution of the amount of genomic elements discovered since 1980 have shown that the start of the HGP in 1990 implied an initial sharp increase in the number of genes discovered (or ‘annotated’, Figure 10), that was suddenly levelled out in the mid-2000s at about 20,000 protein-coding genes, with no change after the publication of the first human genome draft in 2001.⁶⁰ The amount of genetic elements discovered in non-coding regions, in contrast, did experiment an exponential increase with this publication, climbing from 0 in 2000 to 130 thousand before 2020 (Figure 12). These non-coding regions, ignored until that moment to be considered as junk DNA or the dark matter of the genome, revealed to occupy the majority of the human functional sequences (Figure 11); including long non-coding RNAs, promoters, enhancers and other motifs; and to act as gene-regulatory elements of protein expression that work together with the coding regions in a complex and interconnected network to bring the genome to life.

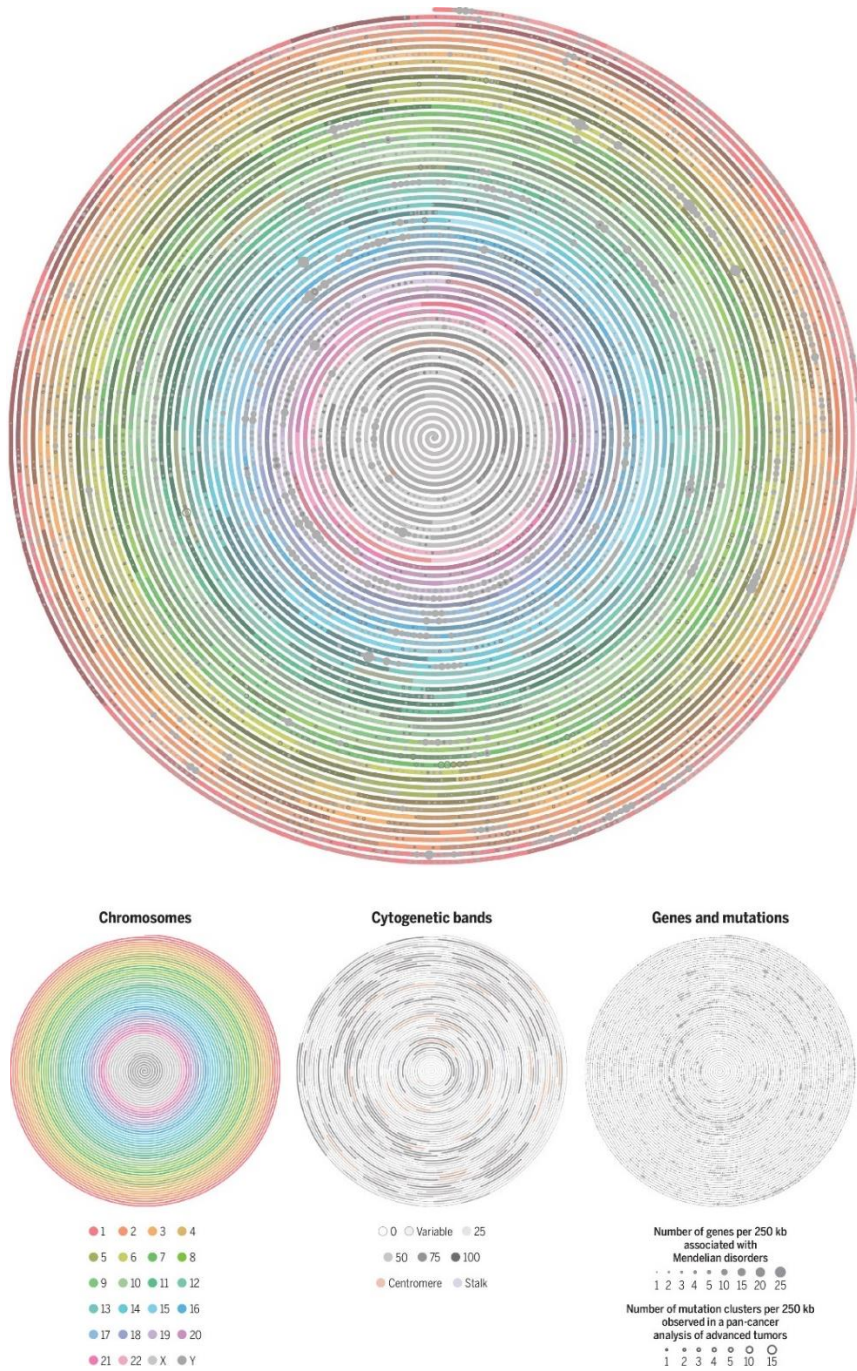


Figure 11. The human genome. (Left) chromosomes 1 through 22 as well as X and Y. (Middle) chromosome features and staining density. (Right) genes associated with Mendelian disorders and genetic mutations found in cancers.⁵⁹

Introduction

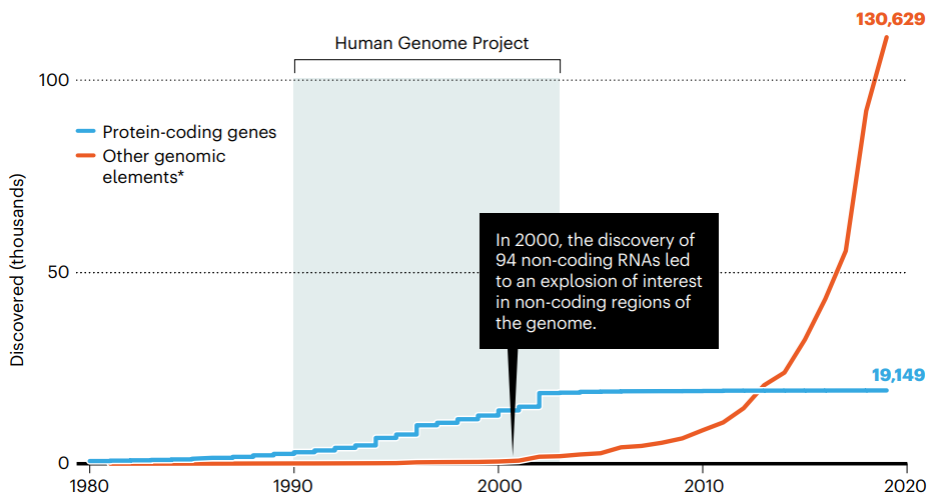


Figure 12. Evolution of the amount of genomic elements discovered in the context of the Human Genome Project. *Including single nucleotide polymorphisms, pseudogenes, non-coding RNAs, promoters and so on. Figure extracted from Gates *et al.* (2021).⁶⁰

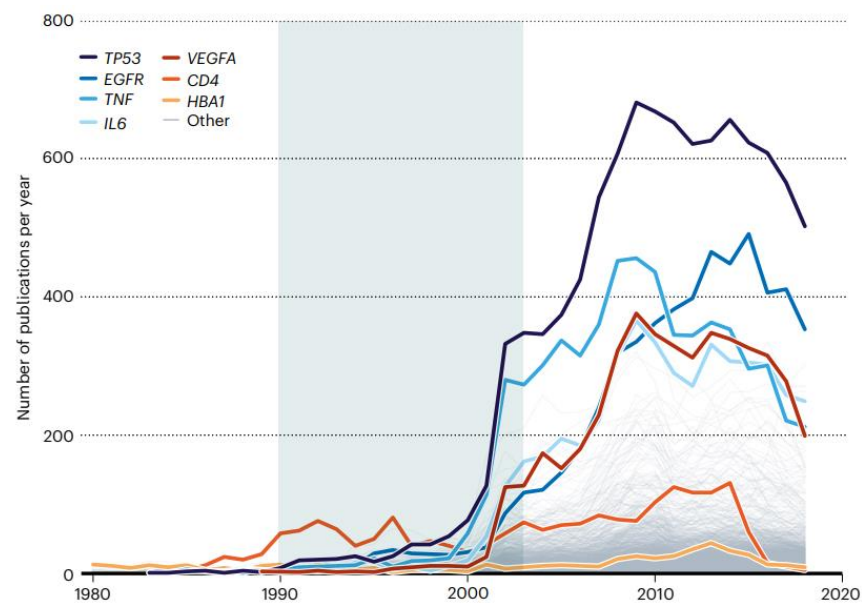


Figure 13. Evolution of the amount of publications associated to genes per year in the context of the Human Genome Project. Figure extracted from Gates *et al.* (2021).⁶⁰

Although discoveries of new protein-coding genes reached a plateau after 2001, the interest in some specific genes grew rapidly following the HGP, most of them associated to cancer settings, such as *TP53*, *EGFR*, *TNF*, *VEGFA* and *IL6* (Figure 13). Each year since 2001, between 10,000 and 20,000 papers were published mentioning these ‘superstar’ genes.⁶⁰ The same trend is observed in the amount of drug targets: of the roughly 20,000 proteins revealed by the HGP as potential drug targets, only about 10% have so far been targeted by approved drugs with for instance 5% of all drugs currently approved (99 distinct molecules) targeting the protein ADRA1A, which is involved in cell growth and proliferation.

There could be good reasons for the detected skew towards specific genes and proteins, like the big importance of some genes in human health or the easier druggability of specific proteins, or it could be that there are many more proteins worth exploring as drug targets if only researchers, funders and publishers were less risk-averse.⁶⁰

I.8 The proteome space

Access to information on gene-coded proteins derived from genome sequencing projects is possible thanks to several open-access repositories such as UniProt, PDB and the very recently launched AlphaFoldDB. **UniProt** (Universal Protein)⁶¹ database was released in 2003 by the UniProt Consortium, which comprises the European Bioinformatics Institute (EBI, <https://www.ebi.ac.uk/>), the Swiss Institute of Bioinformatics (SIB, <https://www.sib.swiss/>), and the Protein Information Resource (PIR, <https://proteininformationresource.org/>). It has since then been maintained with a current content of 0.5 million protein sequences extracted from literature and manually-annotated by curators. For some of these proteins, their 3D structures obtained by X-ray crystallography, NMR spectroscopy or cryo-electron microscopy by biologists and biochemists all around the world are deposited in the **PDB** (Protein Data Bank)⁶² under the supervision of the Worldwide Protein Data Bank (wwPDB). On the other hand, earlier in this year, **AlphaFoldDB** (AlphaFold Protein Structure Database) was launched as a partnership between the EMBL-EBI group and DeepMind company (acquired by Google in 2014).⁶³ In their words, this is the most complete and accurate database yet of predicted 3D protein structures for the human proteome obtained using the artificial intelligence (AI)-based AlphaFold 2 program.⁶⁴ It covers all ~20,000 proteins expressed by the human genome openly available to the scientific community.

Years of study of protein sequences and structures had allowed for the design of classification schemes to organize the known proteome space according to phylogeny. Proteins with common evolutionary origin are grouped in **protein families** including enzymes (protein kinases included), G protein-coupled receptors (GPCRs), ion channels, nuclear receptors,

transporters and transcription factors. Branched from these major families, smaller subfamilies of closer related proteins can be built following a **hierarchical system**⁶⁵ as exemplified in Figure 14. Also, according to their subfamily assignment, proteins can be given multi-level classification codes that summarize their position in the phylogenetic tree and inform about their structure and function. Some examples are the Enzyme Commission Number^{66,67}, used to classify catalytic proteins, and the one stated by the IUPHAR (Union of Basic and Clinical Pharmacology), centred on channels and transporters. Accurate protein classification systems are of vital need in the context of DD for target and hit identification, since, when proper classifications are available, the functional properties and the drug-target interaction information associated to a protein can serve as a highly valuable touchstone to infer the functionality and druggability of their phylogenetically-related yet untargeted proteins, following the paradigm that similar proteins tend to have similar functions and tend to bind similar ligands.^{68,69}

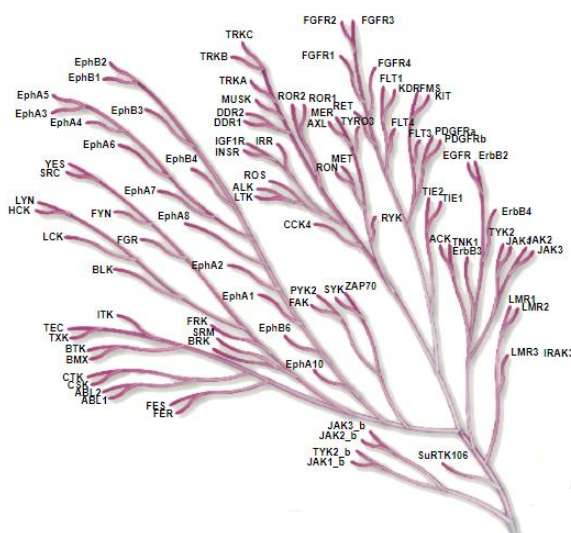


Figure 14. Hierarchical classification of tyrosine kinases subfamilies. Image adapted from <http://www.kinhub.org/kinmap/>.

I.9 The Illuminating the Druggable Genome Initiative

As a complement of the sequence- and structure-based classifications mentioned, in an attempt to classify protein targets according to the level of biomedical and pharmacological knowledge available for them to date and ultimately shed light on areas of understudied ones, the **Illuminating the Druggable Genome** (IDG) initiative, launched by the US National Institutes of Health (NIH) in 2014, proposes the ‘Target Development Level’ classification scheme.⁷⁰ This initiative comprises several American and European universities, hospitals and research centres working together in the IDG consortium. The Target Development Level (TDL) scheme consists of four target categories; namely T_{clin} , T_{chem} , T_{bio} and T_{dark} ; to classify currently known protein targets, with special emphasis in GPCRs, kinases, ion channels and nuclear receptors. T_{clin} (clinic) targets refer to those having an approved drug annotated as its mechanism of action (MoA). T_{chem} (chemistry) proteins stand for those that lack a MoA-based link to approved drugs but that are known to bind to small molecules with high potency. Ligand-target interaction bioactivities for T_{clin} vs T_{chem} discrimination are extracted from ChEMBL and DrugCentral databases already described in Table 1. On the other hand, T_{bio} (biology) refers to those proteins with known biological role and some evidences of linkage to a disease phenotype despite lacking an identified small molecule or approved drug with biological activity for them. Finally, T_{dark} (dark genome) assignments refer to the remaining proteins that have been manually-curated at the primary sequence level in UniProt, yet do not meet any of the criteria for T_{clin} , T_{chem} or T_{bio} . Access to TDL annotations of the druggable genome are publicly available through the **Target Central Resource Database (TCRD)** and **Pharos website**.⁷¹

A snapshot of the distribution of these TDL categories applied to the human proteome in 2018 is shown in Figure 15. From the $\sim 20,000$ targets analysed, only 10% are classified as T_{clin} (3%) or T_{chem} (7%), indicating proteins respectively targeted by approved drugs or other ligands with clinically-relevant activities.⁶⁹ Protein families with higher proportions of clinic and chemistry targets are those having more clinical implications known, such as GPCRs, nuclear receptors or ion channels in central nervous system (CNS) disorders, and kinases in oncological settings. The remaining 90% of the proteome still lacks an identified ligand to target it, including 55% T_{bio} and 35% T_{dark} proteins. Most of these understudied proteins are olfactory GPCRs (oGPCRs), transcription factors, transporters and epigenetic targets, and are the object of study of protein de-orphanization campaigns seeking to elucidate their function and identify bioactive ligands for them.⁶⁸

Introduction

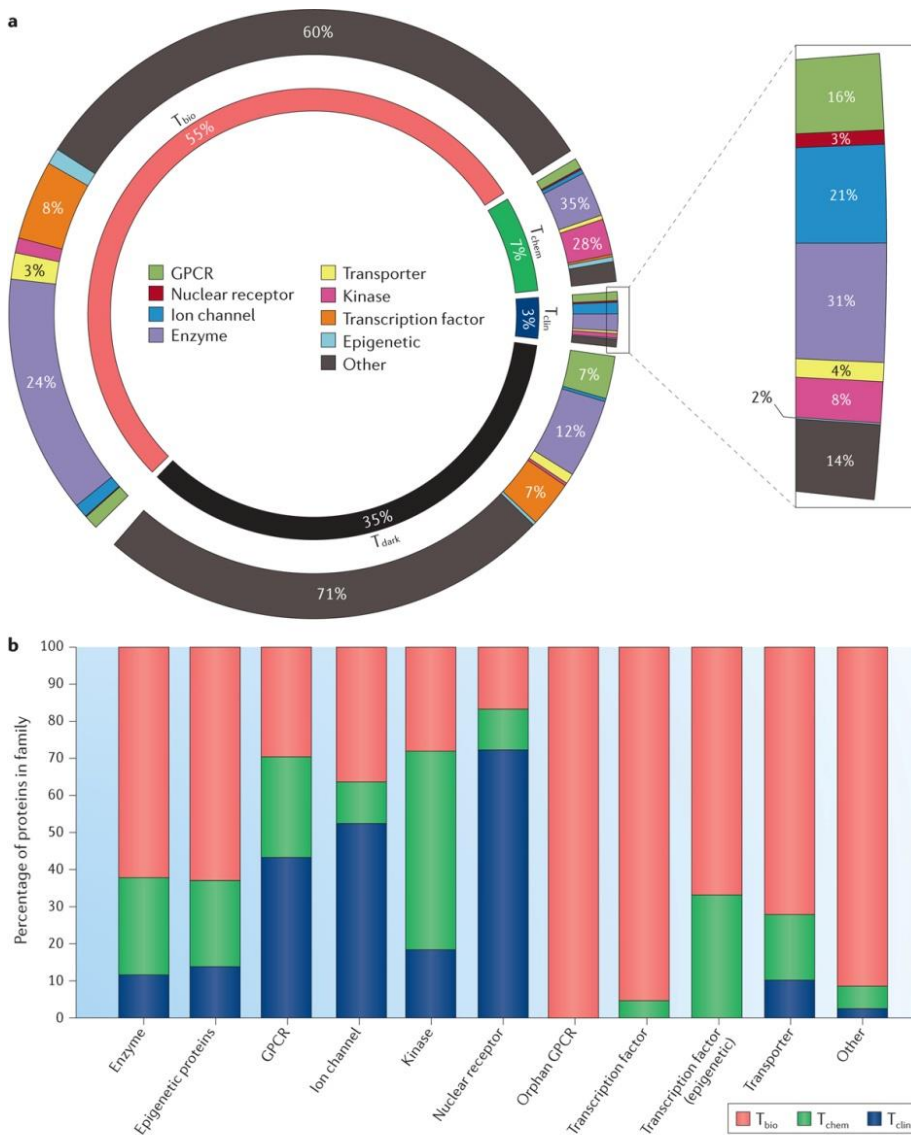


Figure 15. Target development level categories applied to the human proteome. **a)** Percentages of categories and families in the whole proteome. **b)** TDL distribution across protein families. Figure extracted from Oprea *et al.* (2017).⁷⁰

I.10 Polypharmacology in drug discovery

In the previous pages, we have reviewed current knowledge on chemical space and biological space in the context of medicinal chemistry and drug discovery, and have shown that the place where these two spaces collapse is in **Pharmacology**. This is the science that integrates chemistry, biology, pharmacy and medicine to study the molecular mechanisms behind the biological effects that a chemical has in the human body. The advances made in these various fields in the last 20 years, especially those mentioned in this chapter, have substantially improved our understanding on biological systems behaviours and their response to exogenous chemicals and have inevitably cause a necessary shift in the philosophical view that we had of pharmacology.⁷²

In the second half of the 20th century, in a context of genetic reductionism where the ultimate responsible of a specific disease phenotype was believed to be one or two isolated genes,⁷³ finding a drug able to selectively target the individual chemoreceptors derived from those ‘disease causing’ genes and with undesirable off-target effects removed was the standard in drug discovery. This was known as the **‘magic bullet’** paradigm.⁷⁴ After years of focusing on the development of highly selective ligands, accompanied by an unexpected constantly decreasing rate in the amount of drug candidates that were translated into effective clinical therapies in the last decade of the 20th century,⁷⁵ the ‘one gene, one drug, one disease’ thesis was challenged.⁷²

Together with the high clinical attrition rates observed, large-scale functional genomics studies carried out in a variety of model organisms at the beginning of the new century^{76,77,78} evidenced the fact that living cells

Introduction

may not be seen anymore as sets of isolated proteins independently working for a specific phenotype, but as ensembles of complex networks of interconnected proteins, sensible to both intra- and extra-cellular conditions, all together orchestrated to give rise to the different phenotypes. The complexity in this networks revealed redundancy mechanisms of compensatory signalling routes to keep the robustness of the system when one of its elements was perturbed, what may explain the limited efficacy of magic bullet-based drugs.⁷⁹ Thus, in the last years, the classical pharmacology view has slowly been replaced by a ‘network pharmacology’ view, a.k.a **‘polypharmacology’**, which seeks for non-selective or promiscuous multi-target drugs able to perturb not only isolated elements in the cell but whole networks to obtain the desired clinical efficacies⁷² (Figure 16).

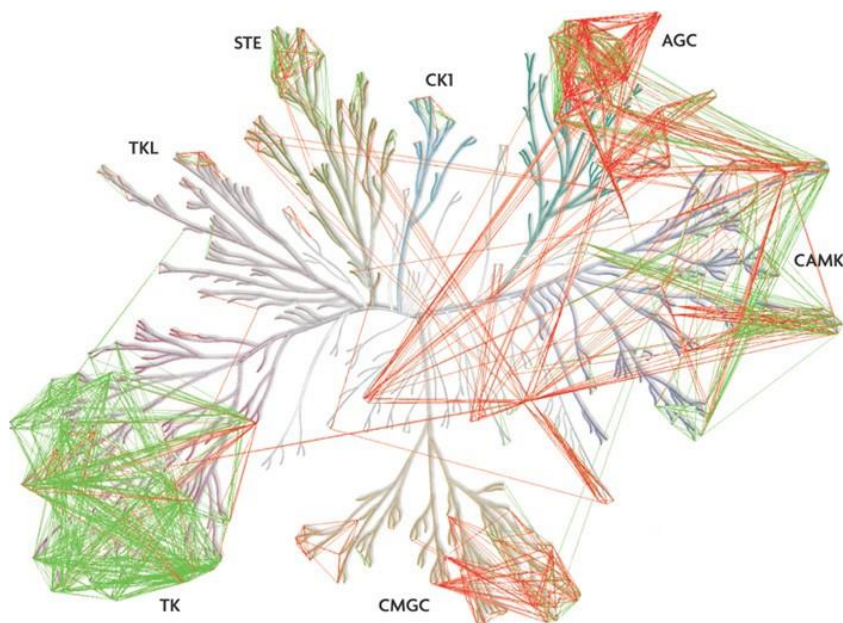


Figure 16. Polypharmacology in the kinases proteome. Pairs of kinases in the tree are connected when they share a common inhibitor. Figure extracted from Knight *et al.* (2010).⁸⁰

In this context, **dual-action drugs** (DADs), defined as compounds that combine two different pharmacological actions at similar efficacious dose,⁸¹ are more and more gaining interest as examples of polypharmacology success stories (Table 3). Applied especially in the treatment of complex diseases; such as cancer,⁸² metabolic disorders⁸³ and CNS disorders,⁸⁴ where single-action drugs' efficacy has seen to be limited, understanding DADs way of action can set the basis for the discovery of other dimensionally-extended poly-action drugs in the future with improved pharmacological profiles.

Table 3. Success stories of dual-action drugs.

Name	Indication	Mechanism of action
Carvedilol ⁸³	Hypertension	Dual beta and alpha-1 adrenoceptor blocker
Bupropion ⁸⁴	Depression	Dual serotonin and dopamine receptors blocker
Bosutinib ⁸⁵	CML*	Dual SRC and ABL inhibitor
Lapatinib ⁸⁶	Breast cancer	Dual EGFR and ERBB2 inhibitor
Clozapine ⁸⁷	Antipsychotic	Dual serotonin and dopamine receptors blocker

*Chronic myeloid leukaemia

It is from this polypharmacological perspective that this Thesis has been conceived.

Part II: Objectives

The main objectives pursued by this Thesis can be summarised as follows:

- i) To design a computational protocol able to identify the core chemical structure best representing the congeneric series of pharmacologically relevant molecules in patents.
- ii) To apply the new methodology to obtain a filtered version of SureChEMBL database enriched with pharmacologically relevant compounds around the patent claim.
- iii) To develop a new similarity-based approach to assess the degree of congenericity of collections of molecules, with emphasis on the claimed molecules from pharmacological patents.
- iv) To design a protocol able to identify those core scaffolds best representing the bioactive chemical series enriched within families of phylogenetically-related proteins.
- v) To apply the new methodology to shed light on the bioactive chemical space of yet untargeted proteins based on their phylogenetically.
- vi) To explore current polypharmacological opportunities for targeting the human proteome by dual-acting agents.

The first and the second objectives were accomplished by the generation of a computationally filtered version of SureChEMBL database, named SureChEMBL_{ccs}, enriched with patent claimed molecules from US pharmacological patents, and available for download at an EMBL ftp site

Objectives

(see **Chapter III.1**). To address the third objective, this newly generated database was used to quantitatively analyse the degree of congenericity of collections of molecules claimed by pharmacological patents (see **Chapter III.2**). The fourth and fifth objectives were accomplished by the identification of a set of family-associated core scaffolds that, when chemically expanded, are highly probable of containing active small molecules for untargeted proteins included in the families (see **Chapter III.3**). Finally, to address the sixth objective, a new ontology of terms to help characterize dual-pharmacological opportunities for targeting the human proteome is proposed and applied to current bioactivity data available in the public domain (see **Chapter III.4**).

Part III: Results

III.1 Identification of the core chemical structure in SureChEMBL patents

Falaguera, M. J. & Mestres, J. Identification of the core chemical structure in SureChEMBL patents. *J Chem Inf Model* **2021**,

61(5), 2241–2247.

Quartile: Q1; Impact Factor 2021: 4.945; Citations: 1

A poster and an oral communication were presented on this topic.

- **Falaguera, M. J.** & Mestres, J. Identification of the core chemical structure in SureChEMBL patents. Poster communication presented at the Symposium to Celebrate 10 Years of the ChEMBL Database. 2019. Hinxton (UK). (See Appendix).
- Falaguera, M. J. Defend your PhD Project in 7 Minutes. Oral communication presented at the PhD Battle at IMIM 1st PhD Day. 2019. Barcelona (Spain).

Identification of the core chemical structure in SureChEMBL patents

Maria J. Falaguera and Jordi Mestres*

JCIM JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

pubs.acs.org/jcim

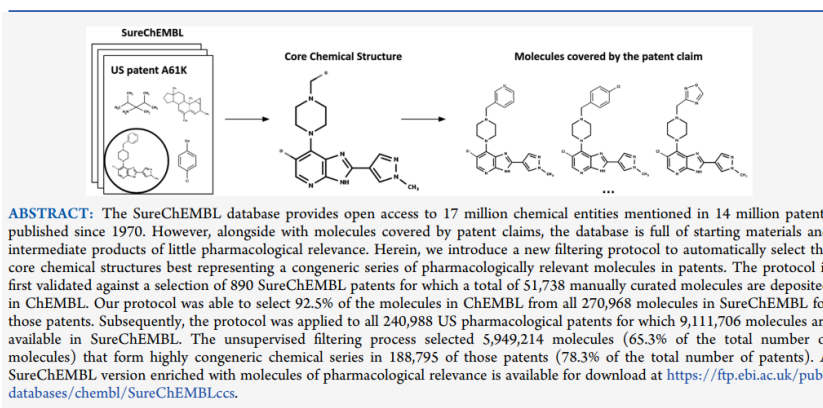
Article

Identification of the Core Chemical Structure in SureChEMBL Patents

Maria J. Falaguera and Jordi Mestres*

Cite This: *J. Chem. Inf. Model.* 2021, 61, 2241–2247

Read Online



ACS Publications

© 2021 American Chemical Society

2241

<https://doi.org/10.1021/acs.jcim.1c00151>
J. Chem. Inf. Model. 2021, 61, 2241–2247

Corresponding Author:

*Email: jmestres@imim.es

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

Introduction

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

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

Results

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

The high patent coverage of SureChEMBL compared to other chemical databases of its kind is the result of applying automated chemical named entity recognition technology to extract every chemical structure from text, images and MOL files attached to the patent document.³ This process ensures the identification and extraction of all chemical entities mentioned in patents. However, this is also one of the recognized limitations of SureChEMBL, as there is no distinction between starting materials, intermediate products and pharmacologically relevant compounds, all ultimately being deposited in the database. To address this situation, Kunimoto and Bajorath⁹ applied the matched molecular pairs (MMP) concept to detect the main substructure shared by the small molecules contained in a patent claim. More recently, Akhondi *et al.* (2019)¹⁰ developed a text-mining recognition system for relevant compounds in a patent based on analyzing the patent's context of a compound defined by its position in the document, the section where it appears, the frequency of appearance, its wide usage in other patents, and any other compounds being mentioned in its textual vicinity. In spite of these efforts, a fully automatic and efficient

process to detect molecules of therapeutic relevance in SureChEMBL patents is still missing.

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

Methods

SureChEMBL database. In the release used in this work (July, 2019), SureChEMBL covered 1,975,722 US patents containing 167,662,929 patent-molecule associations involving 14,284,051 unique small molecules. Out of this total number of US patents, 240,988 (12.2%) can be considered “pharmacological” patents, which contain 45,539,938 patent-molecule associations with 9,111,706 unique small molecules. We define a patent as “pharmacological” when it has an A61K* IPC code, with the exception of A61K6 (preparations for dentistry), A61K7 or A61K8 (cosmetics or similar toilet preparations), A61K9 (medicinal preparations characterized by special physical form), A61K38 (medicinal preparations containing peptides), A61K39 (medicinal preparations containing antigens or antibodies) and A61K48 (medicinal preparations containing genetic material which is

Results

inserted into cells of the living body to treat genetic diseases). However, patents tend to have multiple IPC codes to describe their uses and applications. The most frequent classification code in US A61K* patents is A61K31 (medicinal preparations containing organic active ingredients), but annotations to non-A61K* codes, such as A61P25 (drugs for disorders of the nervous system), C07D401 (heterocyclic compounds containing two or more hetero rings, having nitrogen atoms as the only ring hetero atoms), and C07D413 (heterocyclic compounds containing two or more hetero rings, at least one ring having nitrogen and oxygen atoms as the only ring hetero atoms), are also frequently encountered.

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

Filtering protocol. A filtering protocol was implemented to identify the set of pharmacologically relevant molecules covered by the patent claim from all molecules of a given patent. The protocol is based on the assumption that all relevant compounds in a patent share a core chemical structure that may be represented by an ensemble of candidate maximum common substructures (MCSs) and that these candidate MCSs are

significantly more populated with similar congeneric compounds than any other MCS identified from the other compounds in the patent. The entire process includes three filtering steps and two additional refinement steps, implemented as a concatenated series of Python scripts that are executed sequentially. The main operations were built using the open-source cheminformatics toolkit RDKit¹² version 2017.09.1.

(1) *Extraction of MCSs.* Using as input the SMILES of all compounds in a SureChEMBL patent, the first step is to extract the MCSs for all pairwise combinations of compounds. For this, the `rdkit.Chem.rdFMCS.FindMCS` function is used with the parameters `RingMatchesRingOnly` and `CompleteRingsOnly` activated. A total of 10,377,468 unique MCSs were extracted from all 240,988 US pharmacological patents. At this stage, a promiscuity value, defined as the number of patents in which a given MCS is found, is also assigned to each MCS. About 59% of all unique MCSs are found exclusively in a single patent, whereas less than 3% are present in 10 or more patents (Supplementary Table S1).

(2) *Deletion of promiscuous MCSs.* The main objective of this second step is to discard all molecules in patents likely to be associated with reactants and other substances commonly used in chemistry and thus, unrelated to the patent claims. To this aim, all molecules containing MCSs found above the 1-quantile of the distribution of associated patent promiscuities within a patent were discarded (Supplementary Table S1). About 46% of the patents retained only molecules with MCSs exclusive to them. In contrast, almost 33% of the patents admitted molecules containing MCSs with promiscuities ranging from 1 to 10 or higher.

Results

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

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

(5) *Selection of high confidence patents.* This fifth step is added to assign a confidence label to each patent based on the degree of structural congenericity of the final selected molecules. In this respect, under the assumption that molecules exemplifying a patent claim should define a close congeneric chemical series, the median value of the distribution of pairwise

Dice similarities between all patent molecules will be associated with the level of confidence on the patent. Based on a validation analysis (*vide infra*), patents will be given a “high confidence” flag if the median similarity value is equal or higher than 40%.

Results and discussion

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

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

The second step, involving the removal of molecules associated with promiscuous MCSs, is the one having the strongest filtering effect. A total

Results

of 78,475 molecules are discarded, which correspond to 61.0% of all molecules that will be ultimately filtered out. At this stage, we are left with 192,492 molecules, 71.0% of the initial number of SureChEMBL molecules, which nonetheless include 49,464 molecules present in ChEMBL, that is, 95.6% of all molecules in ChEMBL for those patents.

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

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

Selecting molecules from candidate MCSs only results also in an important reduction of the number of molecules kept from patents. A total

of 50,078 molecules are excluded in this third step, which correspond to 39.0% of all molecules discarded at the end of the three filtering steps. The number of molecules remaining at this stage is 142,414, almost half (52.6%) of the initial number of SureChEMBL molecules. Within them, there are 43,931 molecules present in ChEMBL, which represent 84.9% of all molecules in ChEMBL for those patents.

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

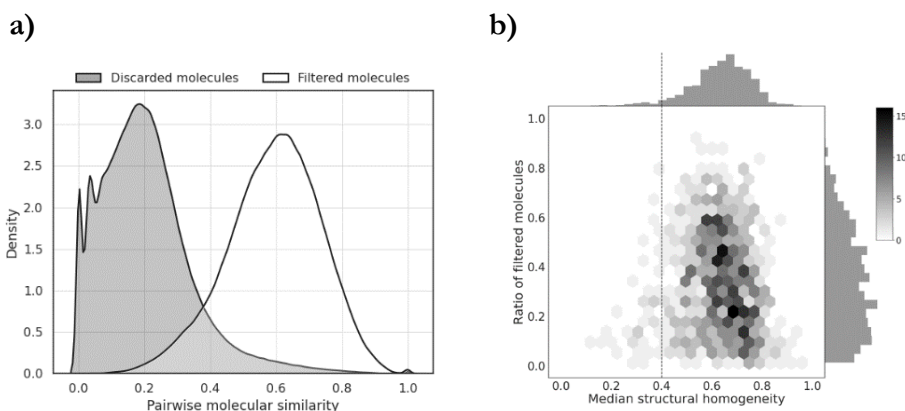


Figure 1. a) Kernel density plot for the distribution of pairwise similarities between the 163,091 filtered molecules (white surface) and the 107,877 discarded molecules (grey surface) up to step 4 of the filtering protocol; **b)** Density plot of median structural homogeneity values against the ratio of filtered molecules in patents. Grey scale of hexagons corresponds to the relative density of patents. Also included are the distributions of the number of patents corresponding to each median structural homogeneity (top x-

Results

axis) and each ratio of filtered molecules (right y-axis). The dotted line at a median structural homogeneity of 0.4 marks the threshold for high confidence patents.

The 163,091 molecules that passed all filters (filtered molecules) within each patent up to this stage should have clearly a higher degree of congenericity than the 107,877 molecules that did not pass any of the filters (discarded molecules). To confirm this assumption, kernel density plots of the pairwise similarity distributions for filtered and discarded molecules were compared (Figure 1a). As it can be observed, there is a clear separation between the two sets, with similarity values at the density peaks of the distributions being 0.60 and 0.19 for filtered and discarded molecules, respectively. A more in-depth analysis would involve adding another dimension to reflect the ratio of filtered molecules remaining in the end within each patent (Figure 1b). As it is shown, most patents have median structural homogeneities between 0.5 and 0.8 and retain between 10% and 60% of the original molecules in SureChEMBL.

A close look at patents having median structural homogeneity values below 0.4 (Figure 1b) revealed that their filtered molecules come from multiple candidate MCSs that may define different regions of a large Markush structure or simply different congeneric series. For these patents, visual inspection of their filtered molecules would be strongly advised. Accordingly, a homogeneity value of 0.4 was established as the lower-bound threshold to identify patents with a high degree of confidence that the final filtered molecules reflect a congeneric series of a well-defined and consistent patent claim. When this threshold was implemented as the last step of the filtering protocol (Table 1), a total of 38 patents were affected, leaving a final number of 851 high-confidence patents, 95.6% of the initial

SureChEMBL patents in ChEMBL. This affected 3,652 molecules in SureChEMBL, leaving the final count of filtered molecules to 159,439, 58.8% of the initial number of SureChEMBL molecules for those patents. This means that over 40% of molecules in those SureChEMBL patents are likely common reactants, intermediate products, and other small molecules not relevant for the patent claims. In contrast, the filtering protocol was able to retain 47,857 molecules present in ChEMBL, that is, 92.5% of all pharmacologically relevant molecules carefully selected and included in ChEMBL from those SureChEMBL patents.

Results for some illustrative examples of high-confidence patents are collected in Table 2. One of them is patent US-8501708 that aims at protecting a class of purine nucleoside compounds as selective A1 adenosine receptor agonists. A total of 115 molecules are present in SureChEMBL. For this particular patent, we found a perfect match between the Markush structure provided in the claim and the only candidate MCS contained in 69 molecules (60% of total) that form a highly congeneric series (median similarity of 0.87). Among them, all 18 molecules (16% of total) contained in ChEMBL were recovered. Another example is patent US-8754099 that protects beta-carboline derivatives as selective antagonists of the somatostatin subtype receptor 3 for the treatment of type-2 diabetes. In this case, SureChEMBL contains 184 molecules, of which 26 (14%) were found to define a highly congeneric series (median similarity of 0.80) around a candidate MCS that matches nicely the Markush structure of the claim. All 4 molecules (2% of total) present in ChEMBL were found within the 26 molecules selected by the filtering protocol. A third example is provided by patent US-8933040 that protects a series of compounds as selective glycosidase inhibitors. Of the 170 molecules in SureChEMBL, 57 molecules (33% of total) were selected by the filtering protocol, among which all 6

Results

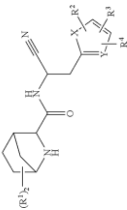
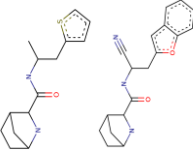

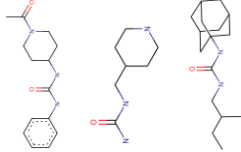
molecules (3% of total) in ChEMBL were present. Note that in this case, not just one but two candidate MCS were identified, both of which match well with the Markush structure of the claim. Another example of a two-candidate MCS case is patent US8871783 that protects the use of 2-azabicyclo[2.2.1]heptane-3-carboxylic acid (cyano-methyl)-amides as cathepsin C inhibitors. A total of 255 molecules were found in SureChEMBL for this patent. Of them, 58 molecules (22% of total) passed all steps of the filtering protocol among which all 26 molecules (10%) present in ChEMBL were recovered. Finally, an example of a case for which the filtering protocol selected three candidate MCS is offered by patent US-8501783. Interestingly, the abstract of the patent states that the invention relates to inhibitors of the soluble epoxide hydrolase that incorporate multiple pharmacophores, therefore justifying the need for multiple candidate MCS to identify all pharmacologically relevant molecules covered by the patent. Of the 190 molecules present in SureChEMBL, 66 molecules (35% of total) passed all filters. In this case, of the 72 molecules contained in ChEMBL for this patent, 51 molecules (71%) were contained within the 66 molecules selected. In general, for any given patent, beyond recovering most of the molecules already in ChEMBL, additional molecules belonging to the same chemical series were retrieved (see details in Supplementary Material). This exemplifies the potential of the approach to produce a version of SureChEMBL containing only molecules around the main core chemical structures (ccs) identified in patents (SureChEMBLccs).

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

Patent ID (median similarity)	Markush structure in patent document	Candidate MCSS selected	Total no. SureChEMBL molecules	Selected no. SureChEMBL molecules (%total)	No. SureChEMBL molecules in ChEMBL (%total sensitivity)
US-8501708 (0.87)			115	69 (60%)	18 (16% 100%)
US-8754099 (0.80)			184	26 (14%)	4 (2% 100%)
US-8933040 (0.65)			170	57 (33%)	6 (3% 100%)

Results

Table 2. (continued)

Patent ID (median similarity)	Markush structure in patent document	Candidate MCSS selected	Total no. SureChEMB L molecules	Selected no. SureChEMB molecules (%total)	No. SureChEMB molecules in ChEMB (%total sensitivity)
US-8871783 (0.64)			255	58 (22%)	26 (10% 100%)
US-8501783 (0.68)			190	66 (35%)	72 (38% 71%)

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

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

Results

27.8% of all molecules that will be filtered out along the process, leaving 8,173,536 molecules at this stage, 89.7% of the initial number of molecules.

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

7,405,130 molecules remained at this stage, 81.3% of the initial number of molecules.

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

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

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

Results

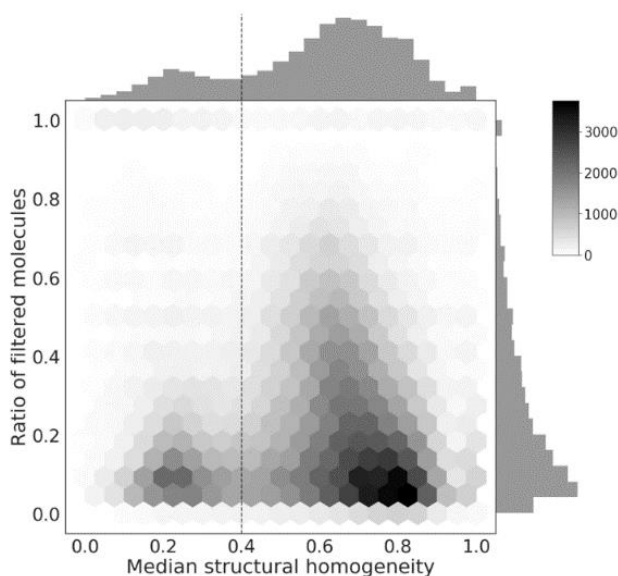


Figure 2. Density plot of median structural homogeneity values against the ratio of filtered molecules in all US pharmacological patents in SureChEMBL. Grey scale of hexagons corresponds to the relative density of patents. Also included are the distributions of the number of patents corresponding to each median structural homogeneity (top x-axis) and each ratio of filtered molecules (right y-axis). The dotted line at a median structural homogeneity of 0.4 marks the threshold for high confidence patents.

Finally, based on the previous validation exercise, a median structural homogeneity threshold of 0.4 was applied to select the list of high confidence patents that contain sets of highly congeneric compounds (Figure 2). The application of this filter affected 47,894 patents, 19.9% of the initial SureChEMBL patents, resulting in the removal of 291,286 molecules. In the end, a total of 5,949,214 molecules were left, 65.3% of all molecules from SureChEMBL US pharmacological patents considered originally.

Conclusion

With the advent of a new generation of artificial intelligence algorithms to recognize and extract chemical structures from patent documents in a more reliable and efficient manner,¹³ unsupervised processes to confidently identify the subset of molecules covered by patent claims from all extracted chemical structures are required. In this work, a filtering protocol was designed to automatically select the core chemical structures best representing a congeneric series of pharmacologically relevant molecules in a patent. To demonstrate the validity of the approach, we applied it first to a set of 270,968 chemical structures from a selection of 890 SureChEMBL patents for which a total of 51,738 manually curated molecules are deposited in ChEMBL. Our protocol was able to identify and discard 41.2% of all molecules in SureChEMBL and retain, within the remaining 58.8%, up to 92.5% of all molecules in ChEMBL. In a second step, we performed a large-scale experiment against 240,988 US pharmacological patents for which 9,111,706 molecules are available in SureChEMBL. With a computational cost of approximately 5 days, our protocol selected 5,949,214 molecules (65.3% of the total number of molecules) that form highly congeneric chemical series in 188,795 of those patents (78.3% of the total number of patents). We believe that this protocol will be useful to assist in the process of producing regular updates of a SureChEMBL version enriched with molecules of pharmacological relevance for the benefit of the entire scientific community.

Funding

This work was supported by a RETOS project from the Spanish Ministerio de Ciencia, Innovación y Universidades (SAF2017-83614-R).

References

1. Papadatos, G. *et al.* SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Res.* **2016**, *44*, D1220–D1228.
2. Senger, S. Assessment of the significance of patent-derived information for the early identification of compound-target interaction hypotheses. *J. Cheminform.* **2017**, *9*, 26.
3. Southan, C. Expanding opportunities for mining bioactive chemistry from patents. *Drug Discov. Today Technol.* **2015**, *14*, 3–9.
4. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930–D940.
5. Heifets, A. & Jurisica, I. SCRIPDB: a portal for easy access to syntheses, chemicals and reactions in patents. *Nucleic Acids Res.* **2012**, *40*, D428–D433.
6. IBM press release, **2011**. <http://www-03.ibm.com/press/us/en/pressrelease/36180.wss> (last accessed on January 28th, 2021).
7. Wang, Y. *et al.* PubChem BioAssay: 2017 update. *Nucleic Acids Res.* **2017**, *45*, D955–D963.
8. Digital Science Ltd. news blog, **2013**. <https://www.digital-science.com/blog/tag/surechem> (last accessed on January 28th, 2021).

9. Kunimoto, R. & Bajorath, J. Exploring sets of molecules from patents and relationships to other active compounds in chemical Space networks. *J. Comput. Aided Mol. Des.* **2017**, *31*, 779–788.
10. Akhondi, S. A. *et al.* Automatic identification of relevant chemical compounds from patents. *A. Database (Oxford)* **2019**, baz001.
11. SureChEMBLccs, **2021**.
<ftp://ftp.ebi.ac.uk/pub/databases/chembl/SureChEMBLccs> (last accessed on January 28th, 2021).
12. Landrum, G. A. RDKit: Open-source cheminformatics software, version 2017.09.1; <http://www.rdkit.org> (last accessed on December 13th, 2021).
13. Staker, J.; Marshall, K.; Abel, R. & McQuaw, C. M. Molecular structure extraction from documents using deep learning. *J. Chem. Inf. Model.* **2019**, *59*, 1017–1029.

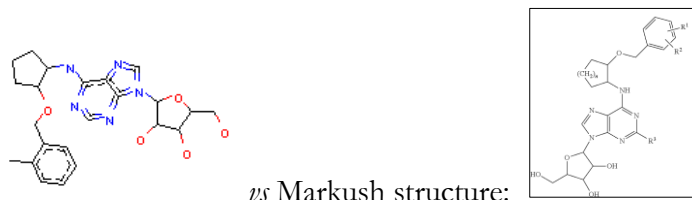
Supplementary material

Table S1. Distribution of the number of MCSs across patent promiscuity values (left) and distribution of the number of patents across 1-quantile threshold values derived from the distribution of patent promiscuities associated with their MCSs (right).

Patent promiscuity	No. MCSs (% from total)	1-quantile promiscuity threshold	No. patents (% from total)
0	0 (0.0%)	0	4,289 (2.0%)
1	6,104,651 (58.8%)	1	111,369 (46.2%)
2	2,242,727 (21.6%)	2	20,450 (8.5%)
3	766,034 (7.4%)	3	8,286 (3.4%)
4	421,330 (4.1%)	4	5,362 (2.2%)
5	213,800 (2.1%)	5	3,647 (1.5%)
6	141,142 (1.4%)	6	2,851 (1.2%)
7	86,151 (0.8%)	7	2,421 (1.0%)
8	62,641 (0.6%)	8	1,954 (0.8%)
9	45,563 (0.4%)	9	1,731 (0.7%)
≥10	293,429 (2.8%)	≥10	78,628 (32.6%)
Total	10,377,468 (100.0%)	Total	240,988 (100.0%)

Patent US-8501708. List of molecules selected by the filtering protocol from all molecules in SureChEMBL.

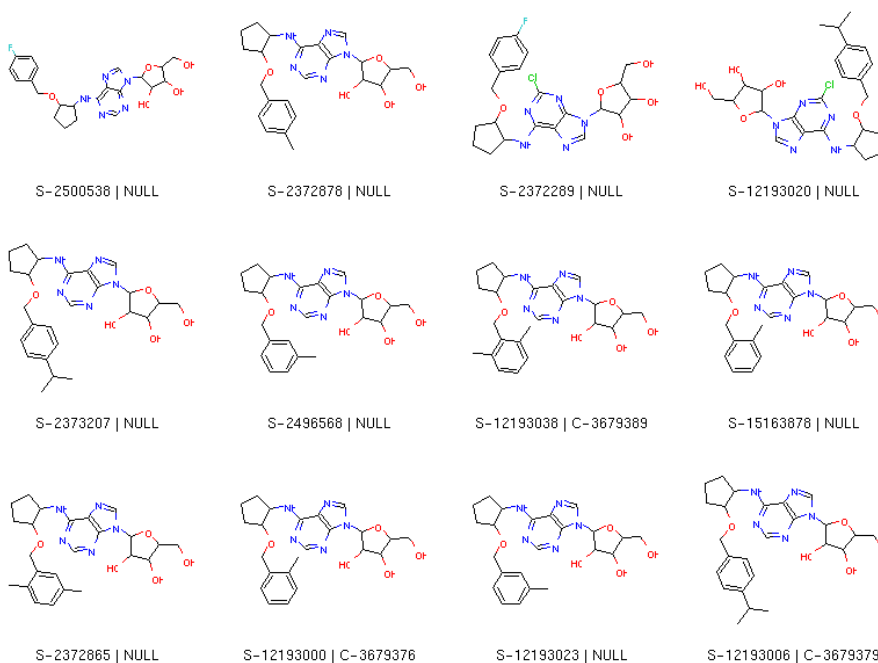
Selected candidate MCS:



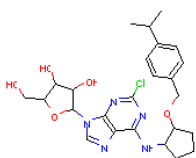
ChEMBL molecules NOT recovered: 0

SureChEMBL molecules selected: 69

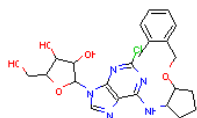
Molecule labels: SCHEMBL ID ('S-*') | ChEMBL ID ('C-*')



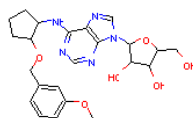
Results



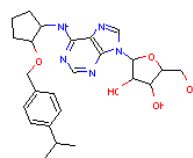
S-2372713 | NULL



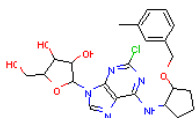
S-2372226 | NULL



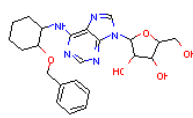
S-12193036 | C-3679388



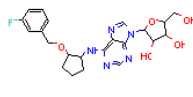
S-12193031 | C-3679386



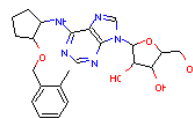
S-2372514 | NULL



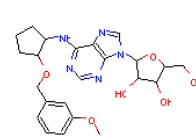
S-15165497 | NULL



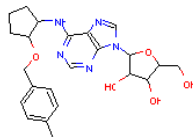
S-12193001 | C-3679377



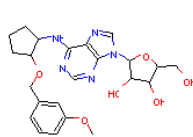
S-12193024 | C-3679383



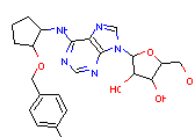
S-12193010 | C-3679381



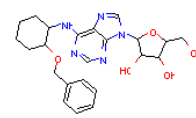
S-2498474 | NULL



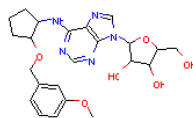
S-2503602 | NULL



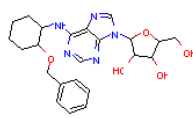
S-12192989 | C-3674590



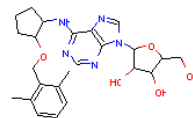
S-12192991 | C-3674591



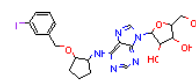
S-2372247 | NULL



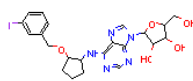
S-2371834 | NULL



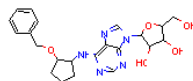
S-2500402 | NULL



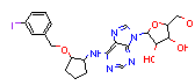
S-2372919 | NULL



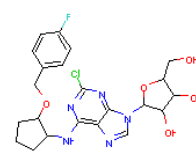
S-2498383 | NULL



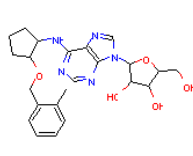
S-2503875 | NULL



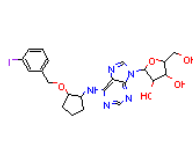
S-15163832 | NULL



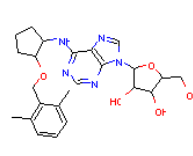
S-12193017 | NULL



S-2504283 | NULL

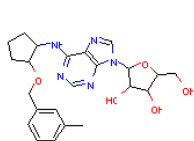


S-12193034 | C-3679387

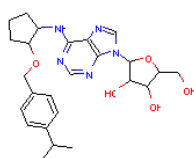


S-2372774 | NULL

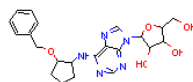
Results



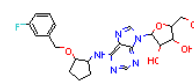
S-2371920 | NULL



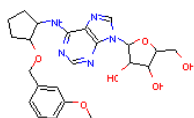
S-2496578 | NULL



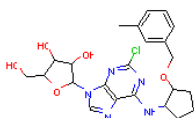
S-12193045 | NULL



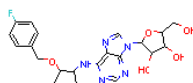
S-2496566 | NULL



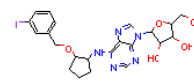
S-2496573 | NULL



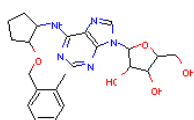
S-15165540 | NULL



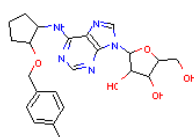
S-12193029 | C-3679385



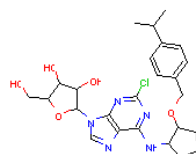
S-2503611 | NULL



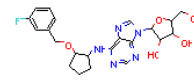
S-2500926 | NULL



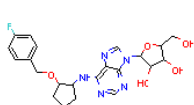
S-2507937 | NULL



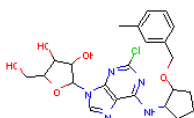
S-15165544 | NULL



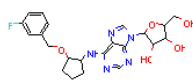
S-12193027 | C-3679384



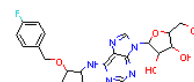
S-2373117 | NULL



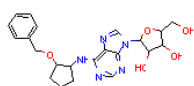
S-12193012 | C-3679382



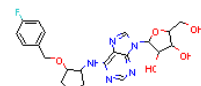
S-2501579 | NULL



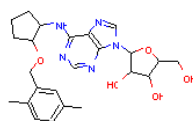
S-2521680 | NULL



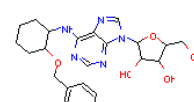
S-2372636 | NULL



S-12193004 | C-3679378



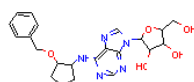
S-12193040 | NULL



S-15165527 | NULL



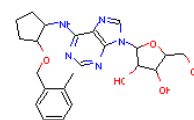
S-15165543 | NULL



S-15165549 | NULL

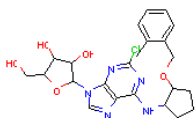


S-12193008 | C-3679380

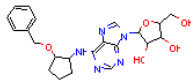


S-2371504 | NULL

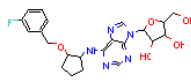
Results



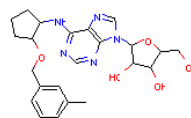
S-12193015 | NULL



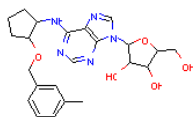
S-12192986 | C-3674589



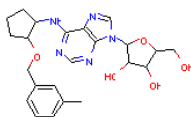
S-2372706 | NULL



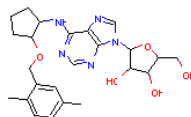
S-2500934 | NULL



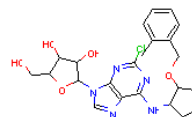
S-12192999 | C-3674592



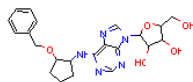
S-15163833 | NULL



S-2507934 | NULL



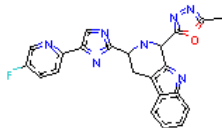
S-15165541 | NULL



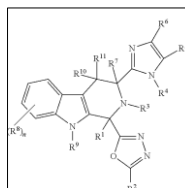
S-15163861 | NULL

Patent US-8754099. List of molecules selected by the filtering protocol from all molecules in SureChEMBL.

Selected candidate MCS:



vs Markush structure:

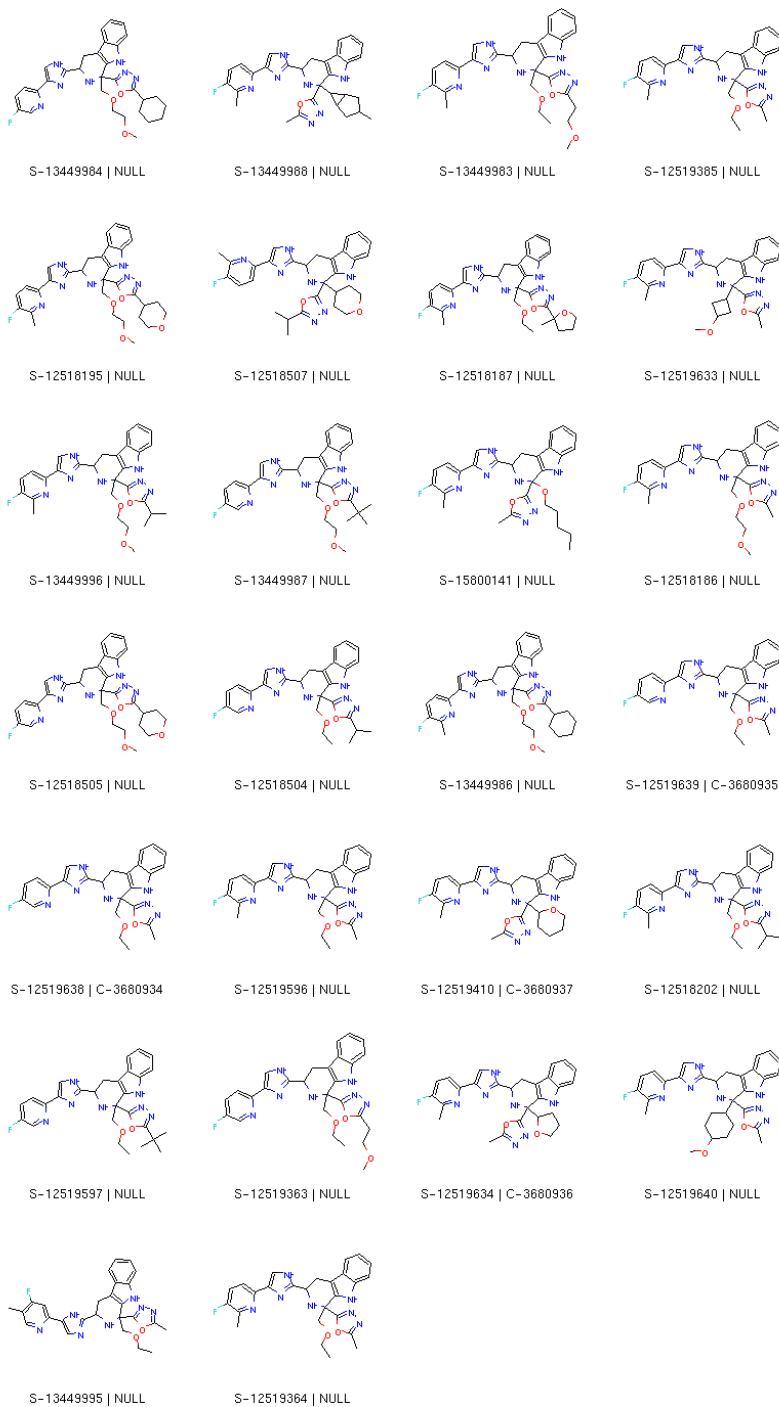


ChEMBL molecules NOT recovered: 0

SureChEMBL molecules selected: 26

Molecule labels: SCHEMBL ID ('S-*') | ChEMBL ID ('C-*')

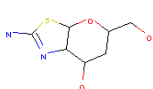
Results



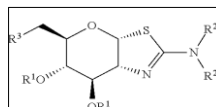
Results

Patent US-8933040. List of molecules selected by the filtering protocol from all molecules in SureChEMBL.

Selected candidate MCSs:



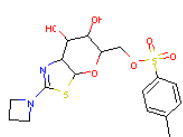
vs Markush structure:



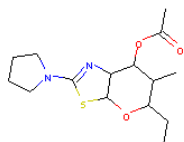
ChEMBL molecules NOT recovered: 0

SureChEMBL molecules selected: 57

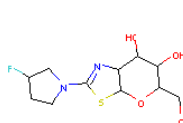
Molecule labels: SCHEMBL ID ('S-*') | ChEMBL ID ('C-*')



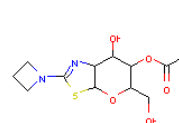
S-4559810 | NULL



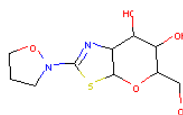
S-15912570 | NULL



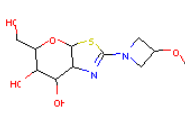
S-15912564 | NULL



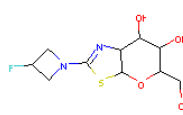
S-15919141 | NULL



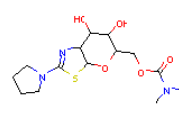
S-4540652 | NULL



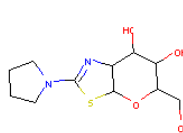
S-15912563 | NULL



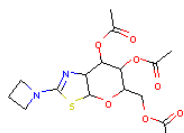
S-15912569 | NULL



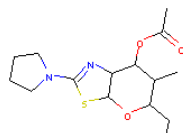
S-4542761 | NULL



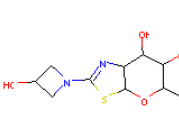
S-4536619 | C-3686728



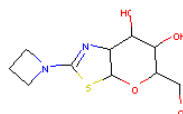
S-5018800 | NULL



S-15912578 | NULL



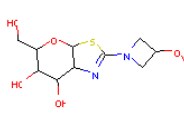
S-7625570 | NULL



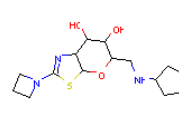
S-4537042 | NULL



S-7621901 | NULL

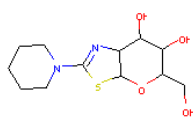


S-4540259 | C-3686727

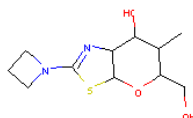


S-4540791 | NULL

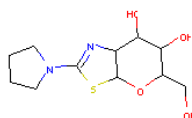
Results



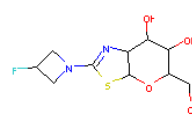
S-4539459 | NULL



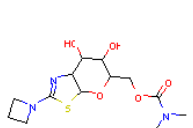
S-15912566 | NULL



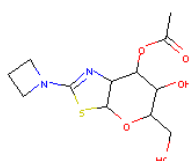
S-14359972 | NULL



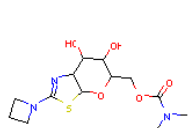
S-4542488 | C-3686726



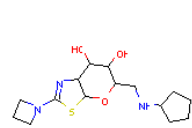
S-7632137 | NULL



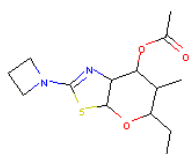
S-15919104 | NULL



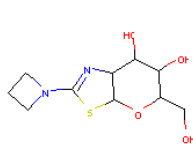
S-4543139 | NULL



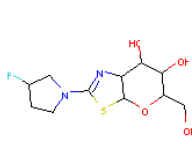
S-15912565 | NULL



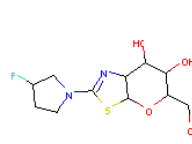
S-15912571 | NULL



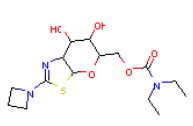
S-14359851 | NULL



S-4542243 | NULL



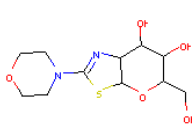
S-15912567 | NULL



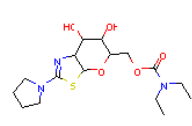
S-5018735 | NULL



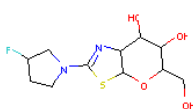
S-15912575 | NULL



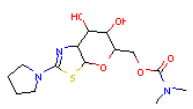
S-4538017 | C-3686729



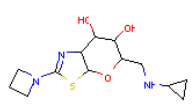
S-7633569 | NULL



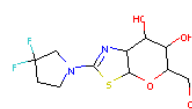
S-15396577 | NULL



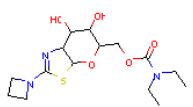
S-5018651 | C-3686731



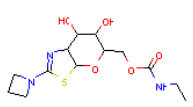
S-4533796 | C-3686730



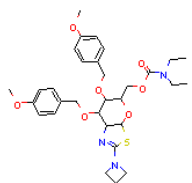
S-15912572 | NULL



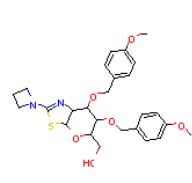
S-4543659 | NULL



S-4538788 | NULL

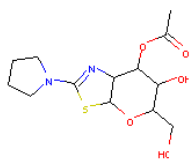


S-4542348 | NULL

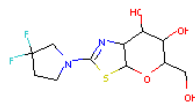


S-4543194 | NULL

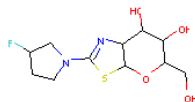
Results



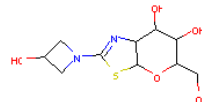
S-15919124 | NULL



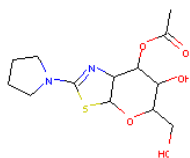
S-4538816 | NULL



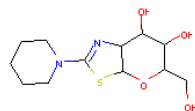
S-4540374 | NULL



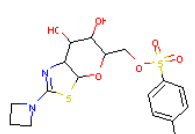
S-4537079 | NULL



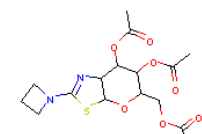
S-15919132 | NULL



S-15912573 | NULL



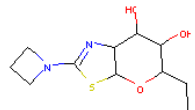
S-4543146 | NULL



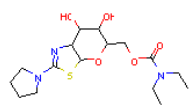
S-15919131 | NULL



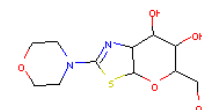
S-5018808 | NULL



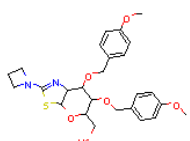
S-4542126 | NULL



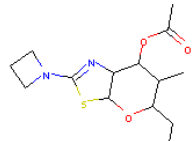
S-4541575 | NULL



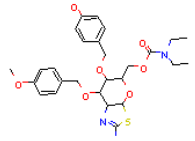
S-15912577 | NULL



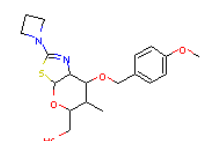
S-5018809 | NULL



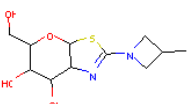
S-15912568 | NULL



S-5018729 | NULL



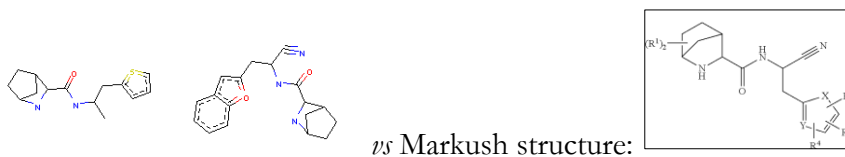
S-15912574 | NULL



S-4542285 | NULL

Patent US-8871783. List of molecules selected by the filtering protocol from all molecules in SureChEMBL.

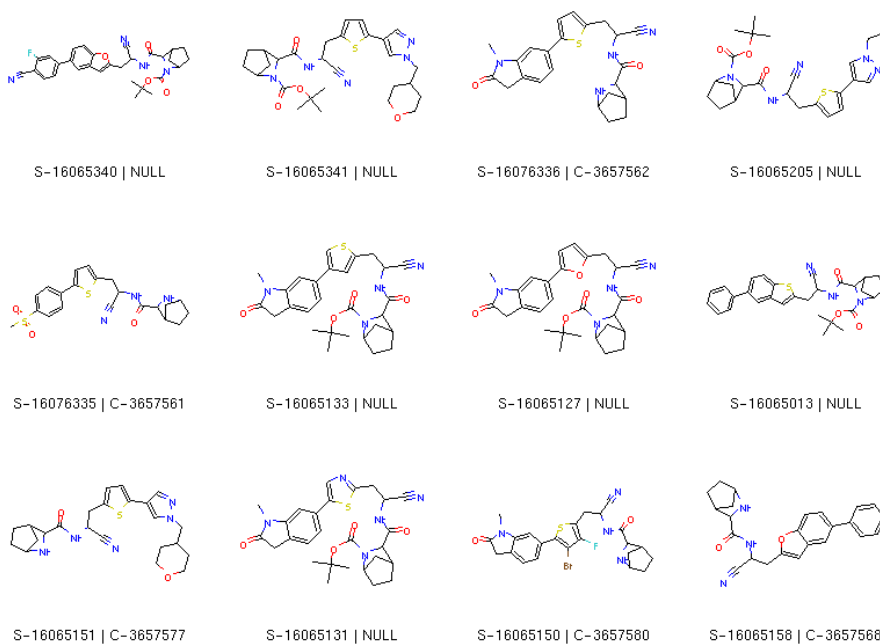
Selected candidate MCSs:



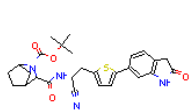
ChEMBL molecules NOT recovered: 0

SureChEMBL molecules selected: 58

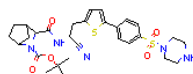
Molecule labels: SCHEMBL ID ('S-*') | ChEMBL ID ('C-*')



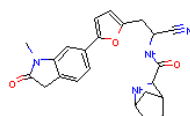
Results



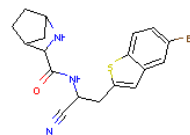
S-16065206 | NULL



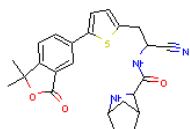
S-16065134 | NULL



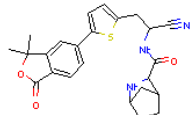
S-16065061 | C-3657564



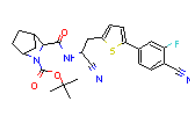
S-16065160 | C-3657569



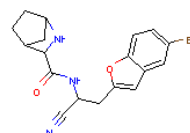
S-16076338 | C-3657583



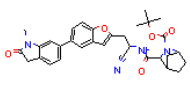
S-16076339 | C-3657584



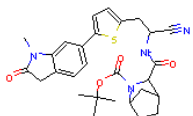
S-16065011 | NULL



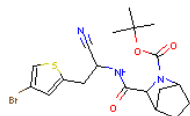
S-16065063 | C-3657565



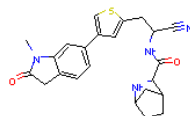
S-16065015 | NULL



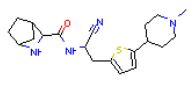
S-16065204 | NULL



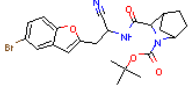
S-16065153 | NULL



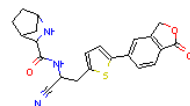
S-16065021 | C-3657574



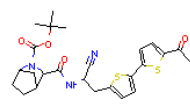
S-16065147 | C-3657579



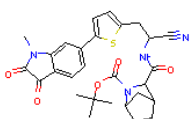
S-16065145 | NULL



S-16065152 | C-3657578



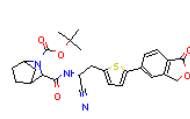
S-16065207 | NULL



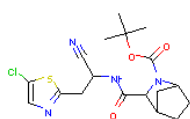
S-16065326 | NULL



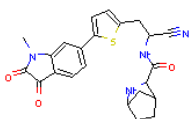
S-16065148 | NULL



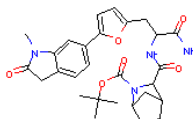
S-16065348 | NULL



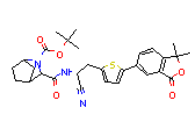
S-16065353 | NULL



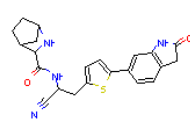
S-16065060 | C-3657563



S-16065126 | NULL



S-16065135 | NULL

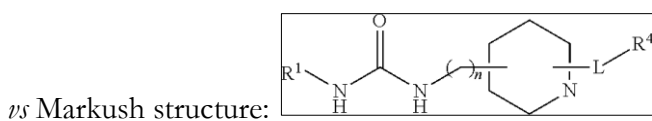
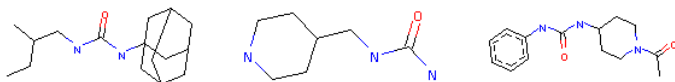


S-16065019 | C-3657576

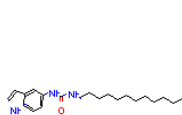
Results

Patent US-8501783. List of molecules selected by the filtering protocol from all molecules in SureChEMBL.

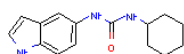
Selected candidate MCSs:



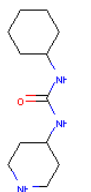
ChEMBL molecules NOT recovered: 21



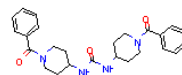
CHEMBL3640215



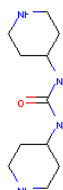
CHEMBL3640216



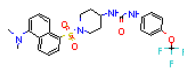
CHEMBL3642238



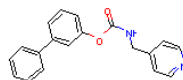
CHEMBL3642240



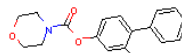
CHEMBL3642239



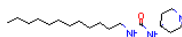
CHEMBL1257880



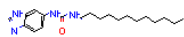
CHEMBL3640217



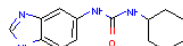
CHEMBL3642234



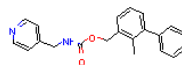
CHEMBL3640213



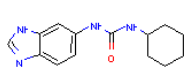
CHEMBL3642235



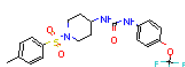
CHEMBL3642236



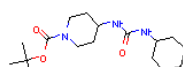
CHEMBL3640218



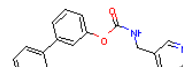
CHEMBL3642236



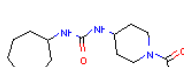
CHEMBL1257759



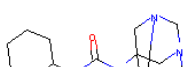
CHEMBL3640214



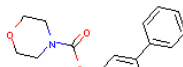
CHEMBL3642231



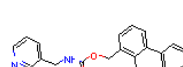
CHEMBL1668928



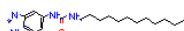
CHEMBL3640212



CHEMBL3642233



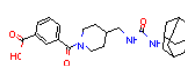
CHEMBL3642232



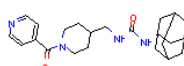
CHEMBL3642235

SureChEMBL molecules selected: 66

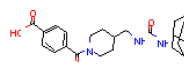
Molecule labels: SCHEMBL ID ("S-*") | CHEMBL ID ("C-*")



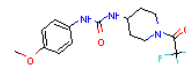
S-5143684 | C-437328



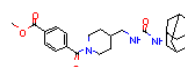
S-4193647 | C-215069



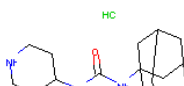
S-5146666 | C-387280



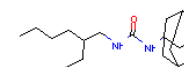
S-14984061 | NULL



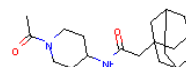
S-4197897 | C-215908



S-4194173 | NULL

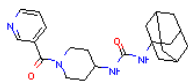


S-15172881 | NULL

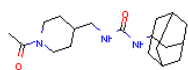


S-4341581 | C-3642242

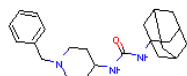
Results



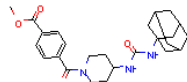
S-5146507 | C-214128



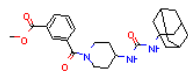
S-652237 | C-215025



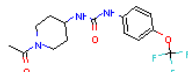
S-4349264 | C-387034



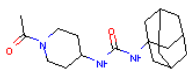
S-4346970 | C-215999



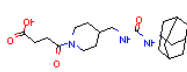
S-5142119 | NULL



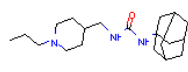
S-5143669 | C-1668934



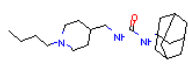
S-654229 | C-436774



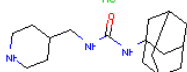
S-5146687 | C-386455



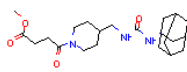
S-8233455 | C-214568



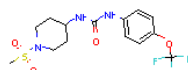
S-5144210 | C-386044



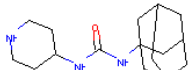
S-5144022 | NULL



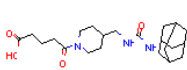
S-4495171 | C-217711



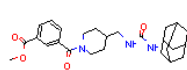
S-1913735 | NULL



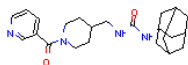
S-8234307 | NULL



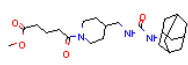
S-5146628 | C-215820



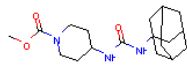
S-8925170 | C-215121



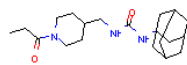
S-5145377 | C-425436



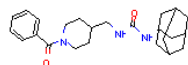
S-4194149 | C-384075



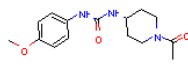
S-4192723 | C-3642245



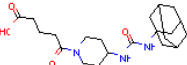
S-5146574 | C-215874



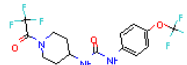
S-5146592 | C-214883



S-14984060 | NULL

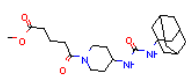


S-4343837 | C-426340

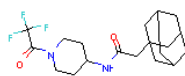


S-2719652 | C-1257517

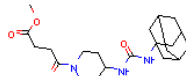
Results



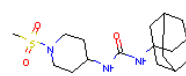
S-4194188 | C-214505



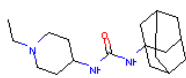
S-5145401 | C-3642244



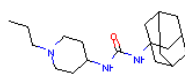
S-5143962 | C-215827



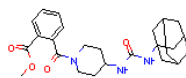
S-3213643 | C-1668935



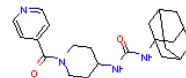
S-5146577 | NULL



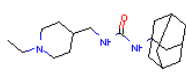
S-5144236 | C-215168



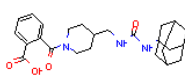
S-5144113 | C-214644



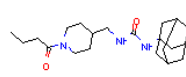
S-4192116 | C-214179



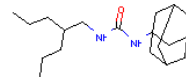
S-8925131 | C-213611



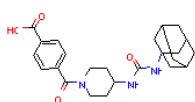
S-4193651 | C-214735



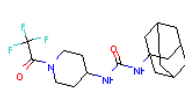
S-4188068 | C-214675



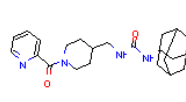
S-15172882 | NULL



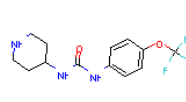
S-4201297 | C-386384



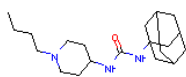
S-853693 | C-217758



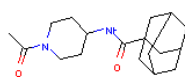
S-5144088 | C-386855



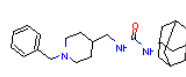
S-2718899 | C-3642237



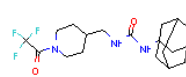
S-5144023 | C-386820



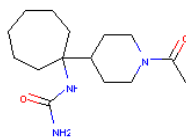
S-4336302 | C-3642241



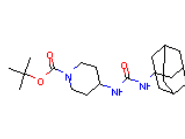
S-4199400 | C-215146



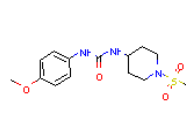
S-8926084 | C-214127



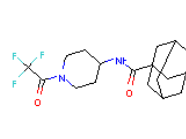
S-8215631 | NULL



S-8925072 | NULL

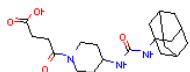


S-14984064 | NULL

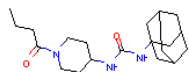


S-8231349 | C-3642243

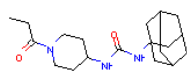
Results



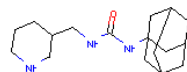
S-5145570 | C-385348



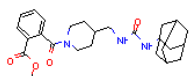
S-4345195 | C-215125



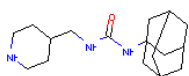
S-2731282 | C-214943



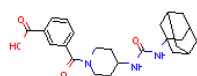
S-15172335 | NULL



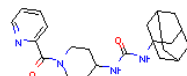
S-4192054 | C-384280



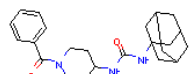
S-8925408 | NULL



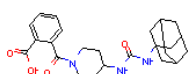
S-4203288 | C-385946



S-4009784 | C-215161



S-4193210 | C-216016



S-4558839 | C-214884

III.2 Congenericity of claimed compounds in patent applications

Falaguera, M. J. & Mestres, J. Congenericity of claimed compounds in patent applications. *Molecules* **2021**, 26(17), 5253.

Quartile: Q1; Impact Factor 2021: 4.148; Citations: 1

The new database generated in Chapter III.1, named SureChEMBLccs, is used in this chapter to quantitatively analyze the degree of congenericity of claimed compounds in pharmacological patents. The results obtained are then compared with those for other patent-derived databases, namely SureChEMBL and ChEMBL.


Congenericity of claimed compounds in patent applications

Maria J. Falaguera and Jordi Mestres*



Article

Congenericity of Claimed Compounds in Patent Applications

Maria J. Falaguera ^{1,2,3} and Jordi Mestres ^{1,2,3,*} 

¹ Research Group on Systems Pharmacology, Research Program on Biomedical Informatics (GRIB), IMIM Hospital del Mar Medical Research Institute, Parc de Recerca Biomèdica (PRBB), Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain

² Department of Experimental and Health Sciences, University Pompeu Fabra, Parc de Recerca Biomèdica (PRBB), Doctor Aiguader 88, 08003 Barcelona, Catalonia, Spain

³ Chemotargets SL, Baldiri Reixac 4, Parc Científic de Barcelona, 08028 Barcelona, Catalonia, Spain

* Correspondence: jmestres@imim.cat

Abstract: A method is presented to analyze quantitatively the degree of congenericity of claimed compounds in patent applications. The approach successfully differentiates patents exemplified with highly congeneric compounds of a structurally compact and well defined chemical series from patents containing a more diverse set of compounds around a more vaguely described patent claim. An application to 750 common patents available in SureChEMBL, SureChEMBLccs and ChEMBL is presented and the congenericity of patent compounds in those different sources discussed.

Keywords: chemical series; patent compounds; similarity analysis; SureChEMBL; SureChEMBLccs; ChEMBL

Molecules **2021**, *26*, 5253. <https://doi.org/10.3390/molecules26175253>

<https://www.mdpi.com/journal/molecules>

Corresponding Author:

*Email: jmestres@imim.es

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

Introduction

A chemical series is a central concept in drug discovery to define a set of small molecules sharing a core structure decorated with different functionalities.¹⁻⁵ These molecular analogues effectively map the chemical space around their common scaffold and thus, they constitute the basis to explore structure-activity relationships⁶ and to identify activity cliffs,⁷⁻⁹ cases of structurally similar compounds with large differences in binding affinities for a given target. When chemical series are enriched with molecules active against several members of a target family their common molecular framework is referred to as a privileged scaffold,¹⁰⁻¹⁵ one of the basic principles in chemogenomics initiatives.¹⁶⁻¹⁸

The development of unsupervised computational protocols for the identification of chemical series in large compound collections has become in recent years an active area of research in chemoinformatics.¹⁹⁻²¹ In particular, applications to automatically detecting the core chemical structure of patent claims among all chemical entities stored in patent databases^{22,23} has received special attention.²⁴⁻²⁶ Because of the use of chemical entity recognition technologies to extract all small molecules from patent text and images, the main difficulty in these cases lies in distinguishing the exemplified compounds intended to be protected from many other starting materials and intermediate products mentioned in the patent. Under the assumption that claimed molecules fit into a chemical series defined by the Markush structure of the patent claim, recent computational protocols have exploited this concept to successfully extract and make publicly available all pharmacologically relevant molecules contained in the largest patent database.²⁷

Results

Having the ability to automatically identify chemical series of claimed compounds in patents, one may then wonder how well the chemical space around the patent claim is covered by those exemplified compounds in the patent. In essence, what we are interested in here is defining a set of quantitative parameters to assess the degree of congenericity of patent compounds. A highly congeneric chemical series of exemplified compounds will be associated with a narrow protection of a well-defined portion of the chemical space, whereas a set of exemplified compounds with low congenericity may reflect a loose attempt to cover the chemical space defined by the patent claim, offering opportunities to fill in the gaps left.

Database and Methods

SureChEMBLccs. We used a subset of the SureChEMBLccs 2021 release²⁷ that includes 159,439 unique small molecules from 851 US pharmacological and high confidence patents²⁶ of which 47,857 molecules are also present in ChEMBL.²⁸ Of those 851 patents, 750 have more than one molecule in ChEMBL. A patent is described as pharmacological when it has an A61K* IPC code, with the exception of A61K6 (preparations for dentistry), A61K7 or A61K8 (cosmetics or similar toilet preparations), A61K9 (medicinal preparations characterised by special physical form), A61K38 (medicinal preparations containing peptides), A61K39 (medicinal preparations containing antigens or antibodies) and A61K48 (medicinal preparations containing genetic material which is inserted into cells of the living body to treat genetic diseases). Also, a patent is considered of high

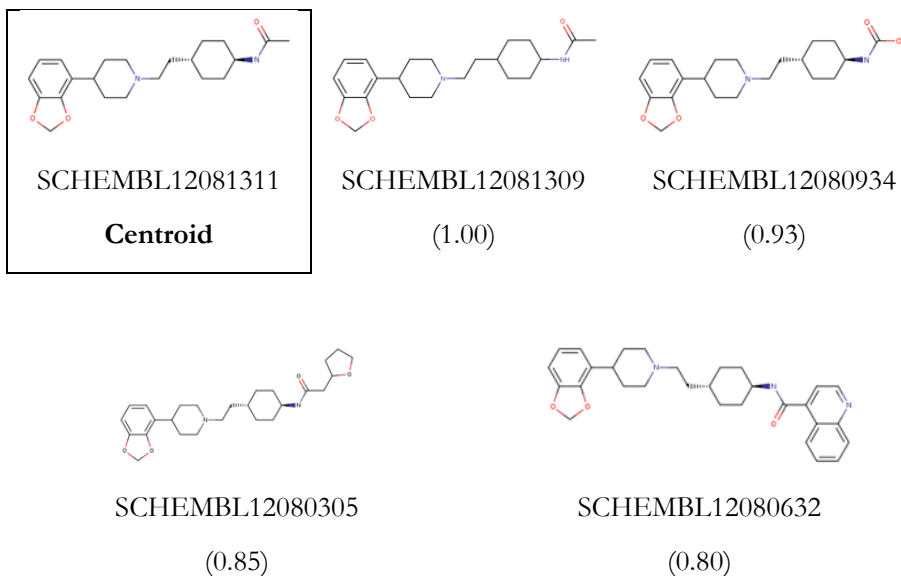
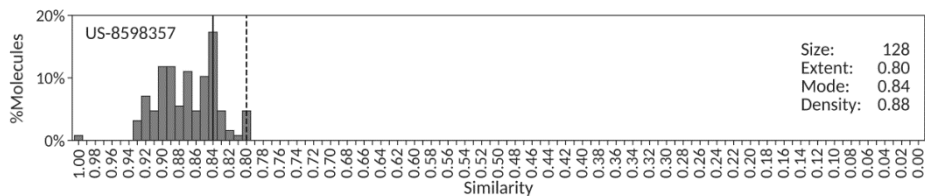
confidence when the molecules exemplified in it describe a similarity distribution with median value equal or higher than 0.4.

Similarity distribution descriptors. A list of three descriptors is suggested to be used in congenericity analyses of sets of molecules. They can be computed from any of the pairwise similarity distributions derived for a given set of N molecules. One intuitive approach is to base the construction of the similarity distribution on the prior selection of a reference molecule. In principle, any molecule can be used as reference. In this work, a *centroid* is selected as the reference molecule having the maximum value of its minimum pairwise similarity. The *centroid similarity distribution* is then defined as the counts of pairwise Dice similarities between the Morgan fingerprints of the centroid against the other $N-1$ molecules in the set, calculated with RDKit,²⁹ within each hundredth of the 0.00 to 1.00 range of similarity values. Alternatively, one can avoid the reference compound selection and directly compute the similarities between all $N*(N-1)/2$ unique pairs of non-identical molecules to construct an *all pairwise similarity distribution*. Then, the shape of the centroid or all pairwise similarity distribution will be quantitatively characterized by its *extent*, *mode*, and *density*. The *extent* is given by the minimum similarity value populated in the distribution. The *mode* is the most frequent similarity value in the distribution. Finally, the *density* measures the degree of dispersion around the *mode*. It is obtained by subtracting from unity the normalized projected Shannon entropy of the distribution,³⁰ that is, the projected optimal number of uniformly occupied bins. A set of compounds having a similarity distribution with high extent, high mode, and high density values will be associated with a highly congeneric series. An additional fourth descriptor, *size*, is considered to account for the number of molecules in the set.

Results and Discussion

Use of similarity distribution descriptors on illustrative patent applications. A set of representative examples were selected to illustrate the use of similarity distribution descriptors to perform congenericity analyses of exemplified compounds in patent applications. The first example is patent US-8598357, a typical case of a patent containing a highly congeneric series of 128 benzodioxole piperidine compounds claimed as dual modulators of the serotonin 2A and dopamine D3 receptors. Figure 1a shows the distribution of pairwise similarity values of all exemplified molecules against the selected centroid, SCHEMBL12081311. The high extent (0.80), high mode (0.84) and high density (0.88) values obtained are all consistent with a highly congeneric chemical series of patent molecules. Correspondingly, the distribution obtained from all 8,128 pairwise similarities (Figure 1b) results also in high extent (0.69), high mode (0.80) and high density (0.81) values. This situation will occur when the core structure common to all exemplified compounds in a patent application covers a large portion of the chemical structures and most molecules differ only by rather small functional groups at one edge of that core structure. This conclusion is confirmed by visual inspection of a selection of molecules with similarity values covering the entire range of the extent in the centroid similarity distribution (Figure 1a).

a) centroid similarity distribution



b) all pairwise similarity distribution

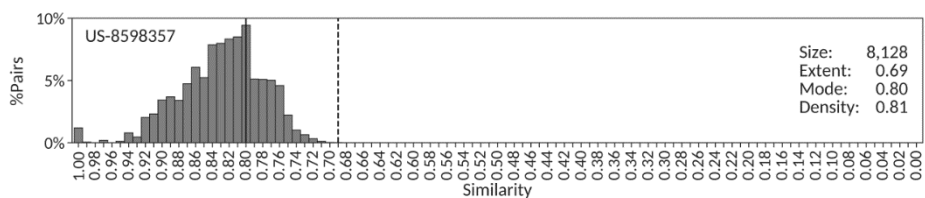
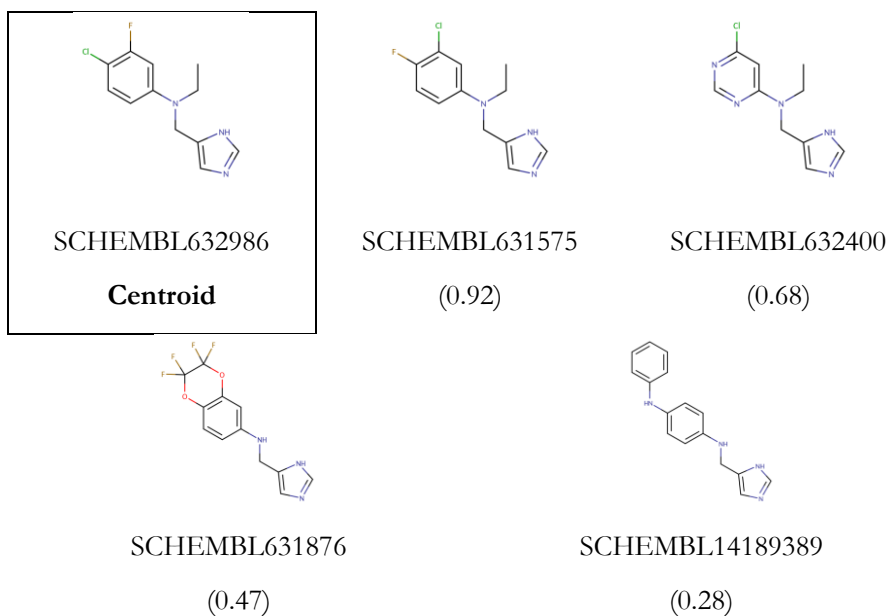
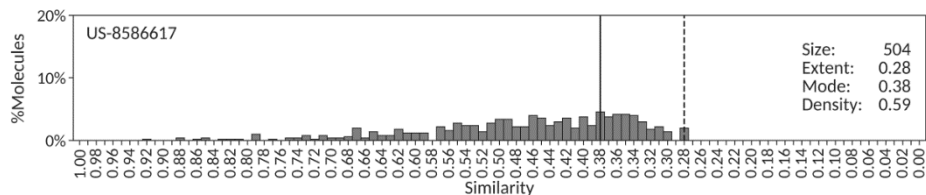


Figure 1. Centroid (a) and all pairwise (b) similarity distributions for the set of exemplified molecules in SureChEMBLccs patent US-8598357. Also included (a) is a sample of chemical structures with their SCHEMBL identifiers and similarity values (in parenthesis). The extent and mode are indicated, respectively, by vertical dashed and solid lines.

Results

A completely different scenario is found in patent US-8586617 protecting a chemical series of 504 amino-4-methyl imidazoles for the treatment of depression, anxiety and bipolar disorders among others. In this case, the centroid similarity distribution around SCHEMBL632986 is quantitatively characterized by low extent (0.28), low mode (0.38) and medium density (0.59) values consistent with a chemical series of patent molecules aiming at sampling diversity rather than coverage completeness (Figure 2a). Similarly, the corresponding distribution derived from all 126,756 pairwise similarities (Figure 2b) results in low extent (0.15), low mode (0.44) and low density (0.49) values. In contrast to the previous patent example, this situation is likely to occur when the common core structure covers only a minor portion of the chemical structure in most of the exemplified compounds in the patent that contains a wide range of diverse, and often large, functionalities around it. This inference is substantiated by the selection of molecules with similarity values covering the entire range of the extent in the centroid similarity distribution (Figure 2a).

a) centroid similarity distribution



b) all pairwise similarity distribution

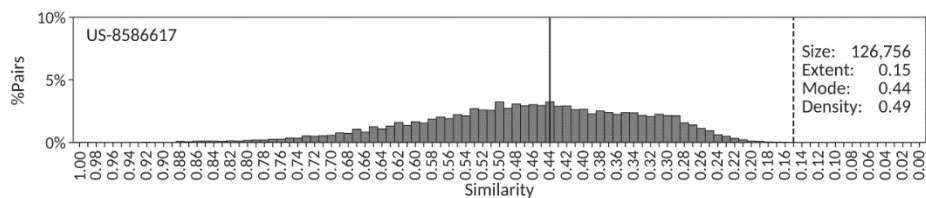
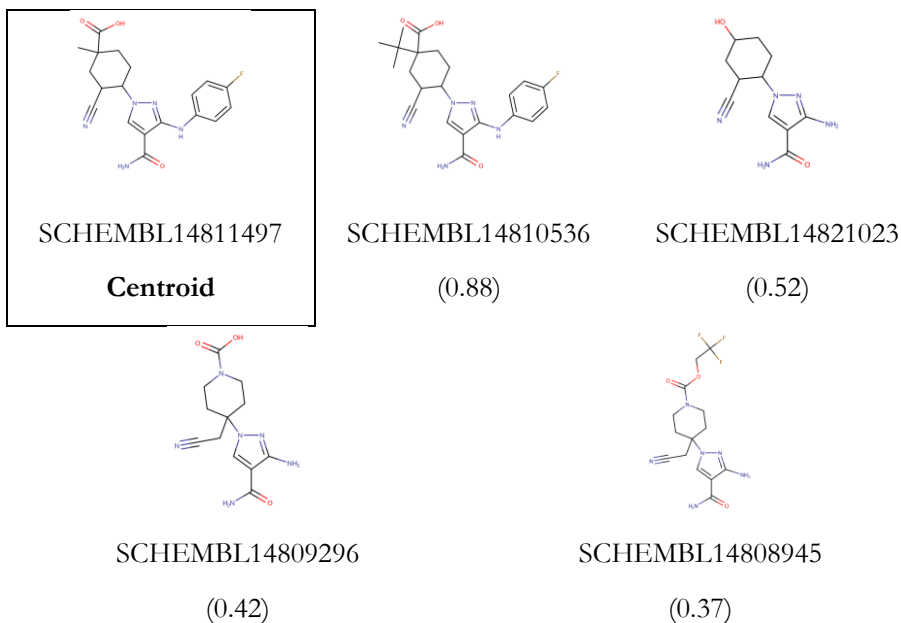
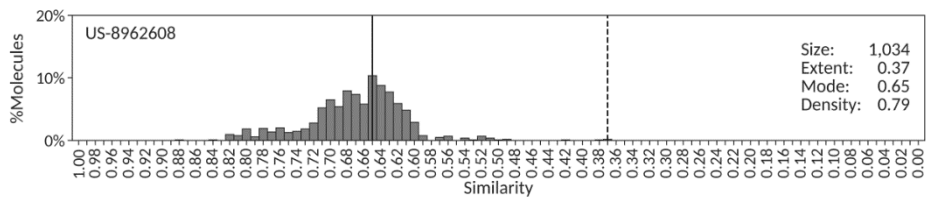


Figure 2. Centroid (a) and all pairwise (b) similarity distributions for the set of molecules exemplified in SureChEMBLccs patent US-8586617. Also included (a) is a sample of chemical structures with their SCHEMBL identifiers and similarity values (in parenthesis). The extent and mode are indicated, respectively, by vertical dashed and solid lines.

Results

An intermediate situation between the two cases presented above is provided by patent US-8962608 aiming at protecting cycloalkylnitrile pyrazole carboxamides as Janus kinase inhibitors. Most of the 1,034 exemplified molecules extracted from this patent show in fact relatively high similarity values (>0.60) against the centroid, SCHEMBL14811497, but a small number of molecules form a long tail below that similarity mark with values as low as 0.37 (Figure 3a). Therefore, despite the medium mode (0.65) and high density (0.79) values, the centroid similarity distribution has also a low extent (0.37) value. The corresponding distribution derived from all 534,061 pairwise similarities (Figure 3b) follows very much the same trend, with a low extent (0.26) value despite the relatively high mode (0.67) and density (0.67) values. This may help in the unsupervised identification of patents for which a large subset of the exemplified compounds extracted automatically form a reasonably tight congeneric series but this congenericity is somehow masked with medium to low extent values due to the presence of a few distant compounds that nonetheless share some core structure attributes. As can be observed in Figure 3a, this is indeed the case for this patent because some intermediate products were recognized by the automatic extracting protocol²⁶ as being part of the core structure of claimed compounds (SCHEMBL14821023, SCHEMBL14809296 and SCHEMBL14808945).

a) centroid similarity distribution



b) all pairwise similarity distribution

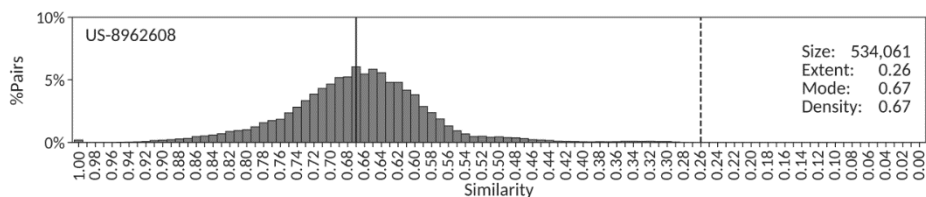


Figure 3. Centroid (a) and all pairwise (b) similarity distributions for the set of molecules exemplified in SureChEMBLccs patent US-8962608. Also included (a) is a sample of chemical structures with their SCHEMBL identifiers and similarity values (in parenthesis). The extent and mode are indicated, respectively, by vertical dashed and solid lines.

Results

The shape of similarity distributions and the descriptor values derived from them depend ultimately on the choice of the reference compound. In the patent examples presented above, the similarity centroid was selected as reference compound. To assess the dependency of the similarity distribution descriptors on the selected reference, a random sample of 10% of all compounds from each patent was extracted, each compound selected as individual reference, and the corresponding similarity distributions and associated descriptors calculated. The results reveal that, even though the exact values of extent, mode and density may vary slightly, the overall quantitative description of the different similarity distributions obtained from random claimed compounds in a patent is essentially retained. Accordingly, the corresponding mean and standard deviation values for extent, mode and density using 13 random compounds as references to derive the similarity distributions of patent US-8598357 are 0.74 ± 0.02 , 0.82 ± 0.02 and 0.86 ± 0.02 , respectively, not too distant from the values reported in Figure 1a. Similarly, the extent, mode and density values calculated from the similarity distributions constructed when using a set of 50 random claimed compounds from patent US-8586617 are 0.21 ± 0.03 , 0.46 ± 0.10 and 0.61 ± 0.06 , respectively, all close to the values shown in Figure 2a, and those resulting from taking 103 random compounds from patent US-8962608 are 0.31 ± 0.02 , 0.65 ± 0.07 and 0.76 ± 0.04 , respectively, all values near those reported in Figure 3a. Therefore, even though similarity distribution descriptors depend on the reference compound selected, the variability observed in their exact values does not affect the ability of the descriptors to capture quantitatively the essence of the degree of congenericity in sets of claimed patent compounds.

The alternative to using a reference molecule to construct the similarity distribution is to simply account for all pairwise similarities. The advantages

are that it alleviates the reference compound selection dilemma, and it provides a unique, more robust, similarity distribution. However, there are also some disadvantages worth considering. For example, the selection of a centroid from which the similarity distribution is derived offers a sense of chemical space coverage around a molecular structure central to the set of patent molecules that cannot be obtained from an all pairwise similarity distribution. Also, one may argue that plotting the percentage of molecular pairs instead of the percentage of molecules in the corresponding similarity distributions provides a less intuitive picture of the similarities between molecules and may confound comparisons between patents. Balancing all these advantages and disadvantages and considering also the good correspondence between descriptor values obtained from centroid and all pairwise similarity distributions observed in the three patent examples presented above, centroid similarity distributions will be used in the remainder of this work.

Correlations of similarity distribution descriptors across patents.

Having illustrated the use of similarity distribution descriptors to quantify the degree of congenericity of claimed compounds in three patent examples, the set of 851 US pharmacological and high confidence SureChEMBLccs patents present also in ChEMBL was processed. In terms of size, the number of claimed molecules per patent ranged from 2 to 2,790, with a median value of 98 molecules, and 425 (50%) and 785 (92%) of the patents containing more than 100 and less than 600 molecules, respectively. The analysis of the centroid similarity distributions gave a wide range of extent values, ranging from 0.09 to 1.00, with a median value of 0.50, and of mode values from 0.12 to 0.96, with a median value of 0.67. It is worth mentioning here that the limit case of extent values equal to unity is due to seven patents

Results

(e.g. US-8895245) having two molecules (e.g. SCHEMBL804176 and SCHEMBL803938) with identical Morgan fingerprints but different structures. In contrast, density values tended to be relatively high for all 851 patents, with minimum and median values of 0.53 and 0.78, respectively. Overall, the median values obtained for the three similarity distribution descriptors reflect the fact that patent molecules in SureChEMBLccs form rather compact chemical series around common core chemical structures and provide further reassurance of the filtering protocol applied to extract them from the original SureChEMBL database.²⁶

Examination of the potential existence of pairwise correlations between the descriptors obtained for the 851 patents resulted in the identification of both positive and negative correspondences. As shown in Figure 4, the strongest correlation identified ($r^2 = +0.70$) is between extent and mode, a trend exposing that, on one hand, patents having high extent values necessarily accumulate pairs of molecules at high mode values (see Figure 1) and, on the other hand, as extent values decrease, the similarity distribution tends to disperse its bin population across the extent and thus, the mode values have also a tendency to decrease accordingly. In this respect, almost 79% of the patents (669) have minimum similarity values (extent) within 0.25 orders of magnitude from the mode indicating that patents with high modes and low extents are exceptional. The second strongest correlation, albeit negative ($r^2 = -0.68$), is between density and size. This is an expected situation as the larger the number of claimed molecules in a patent, the larger its chemical diversity in principle is and thus, the more difficult that pairwise similarities accumulate around the mode, resulting in lower density values. The third trend encountered ($r^2 = +0.52$) is between extent and density, which is consistent with the fact that molecular sets having high minimum pairwise similarities (high extent) are

more likely to have their similarity distributions concentrated in a small number of bins (high density) and vice versa. No significant relationships were found between extent and size, mode and size and mode and density.

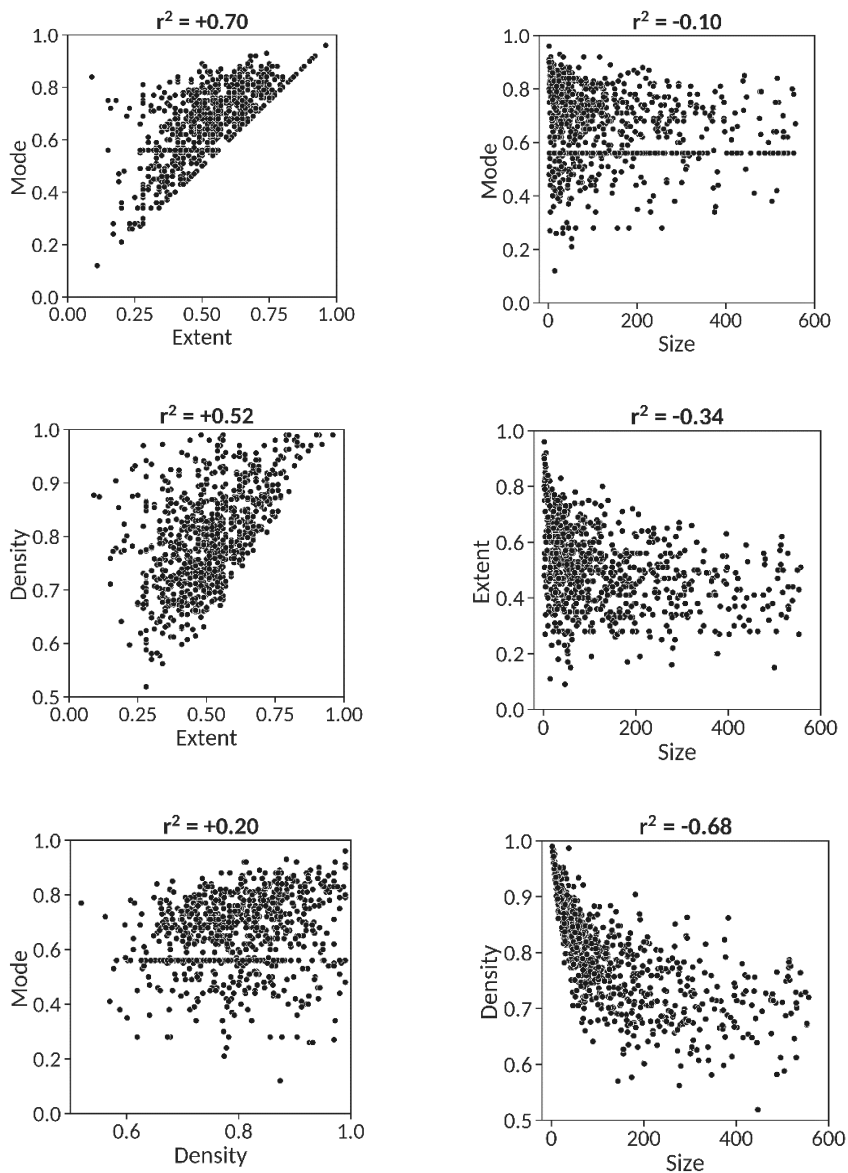


Figure 4. Pairwise correlations between the four descriptors obtained from centroid similarity distributions of 851 patents. The correlation coefficient (r^2) is provided on top of each graph.

Results

To illustrate these correlations with concrete examples, a set of four patents was selected with consistent high density values (0.73) and varying extent, mode and size values. Their corresponding centroid similarity distributions are shown in Figure 5. The first patent (US-8999998) contains a highly congeneric chemical series of 550 pyrazolopyrimidine Janus kinase inhibitors, with a centroid similarity distribution characterized by a medium extent (0.50) and a high mode (0.80). The second patent (US-8637507) is composed of 155 heterocyclic compounds as diacylglycerol acyltransferase inhibitors that has a centroid similarity distribution of comparable density to US-8999998 but with slightly higher extent (0.56) and lower mode (0.71). The third (US-8815891) and fourth (US-9073870) patents exemplify, respectively, 310 tricyclic derivatives as poly(ADP-ribose) polymerase inhibitors and 464 alicyclic carboxylic acid derivatives of benzomorphan and related scaffolds as 11 β -hydroxysteroid dehydrogenase 1 inhibitors. Despite having consistent density values with the first two patents, their centroid similarity distributions have clearly lower extent (0.34 and 0.28, respectively) and lower mode (0.50 and 0.41, respectively) values consistent with sets of more diverse compounds that nonetheless share a core chemical structure.²⁶

The shape of the four centroid similarity distributions shown in Figure 5 is representative of the average similarity distribution obtained for the set of 851 SureChEMBLccs patents analyzed in this work, with average density values of 0.78 (*vide supra*). Comparing the values of the descriptors across the four patents, the positive trend between extent and mode detected above (Figure 4) is recovered and can be visually assessed. For the remaining descriptor pairs, it becomes evident that no clear trend can be established.

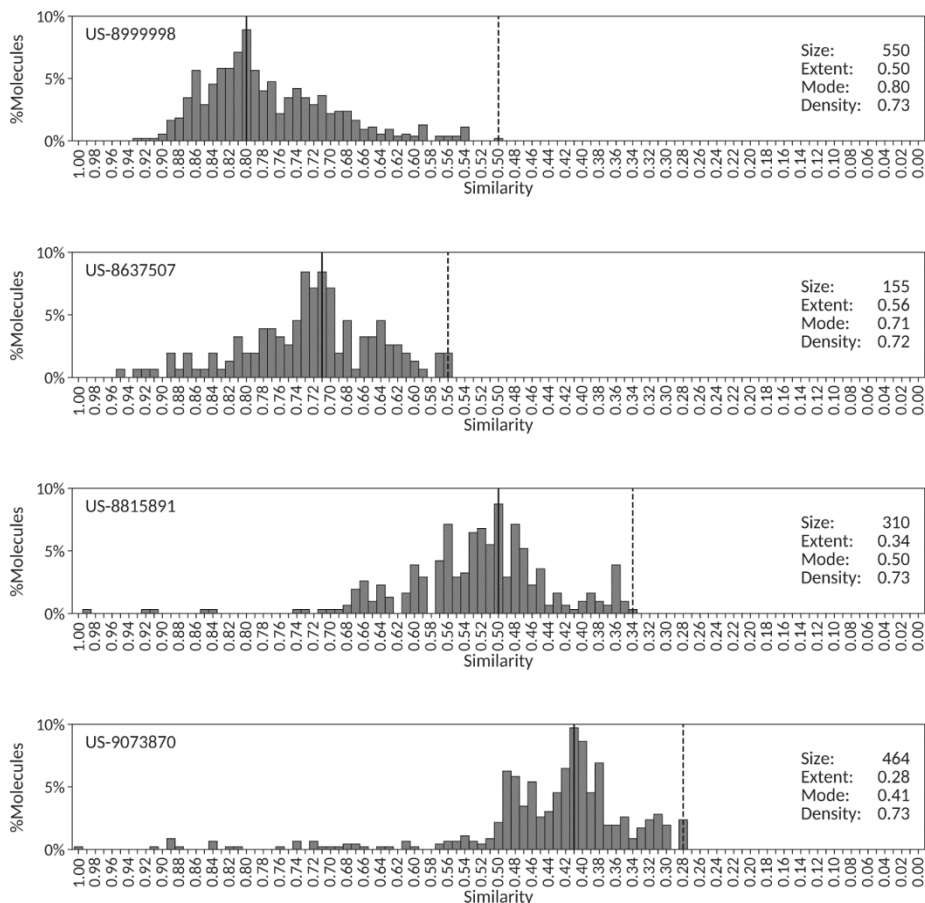


Figure 5. Centroid similarity distributions for four patents with almost identical density values. The mode and the extent are indicated, respectively, by vertical solid and dashed lines.

As a final remark, it is worth stressing that the level of precision in the wording of the patent summary defining the chemical nature of the compounds being claimed already provides some clues on the expected degree of congenicity for the set of exemplified compounds in those patents. For example, defining a set of pyrazolopyrimidine inhibitors in patent US-8999998 is a more chemically precise wording that the generic

Results

mention of alicyclic carboxylic acid derivatives of benzomorphans and related scaffolds in patent US-9073870, and this is then clearly reflected in the differences between extent (0.50 *vs* 0.28) and mode (0.80 *vs* 0.41) values. This aspect could be exploited in the use of text-mining techniques when processing patent titles and summaries.

Congenericity analysis of SureChEMBL patents. SureChEMBL_{ccs}²⁷ was derived by applying an unsupervised automatic filtering protocol to identify the core chemical structure in SureChEMBL patents and extract all pharmacologically relevant molecules exemplifying the patent claims.²⁶ Accordingly, SureChEMBL_{ccs} should be in principle intrinsically biased towards highly congeneric chemical series of compounds. In order to assess this assumption and validate the use of similarity distribution descriptors to quantify congenericity in sets of molecules, a principal component analysis (PCA) was performed on a focused set of 750 SureChEMBL patents for which a filtered subset of compounds sharing a core chemical structure is available in SureChEMBL_{ccs} and a carefully curated selection of at least two compounds is also present in ChEMBL. For the PCA, each patent was quantitatively defined by the extent, mode and density values of the centroid similarity distributions derived from the corresponding full set of compounds in SureChEMBL, SureChEMBL_{ccs}, and ChEMBL.

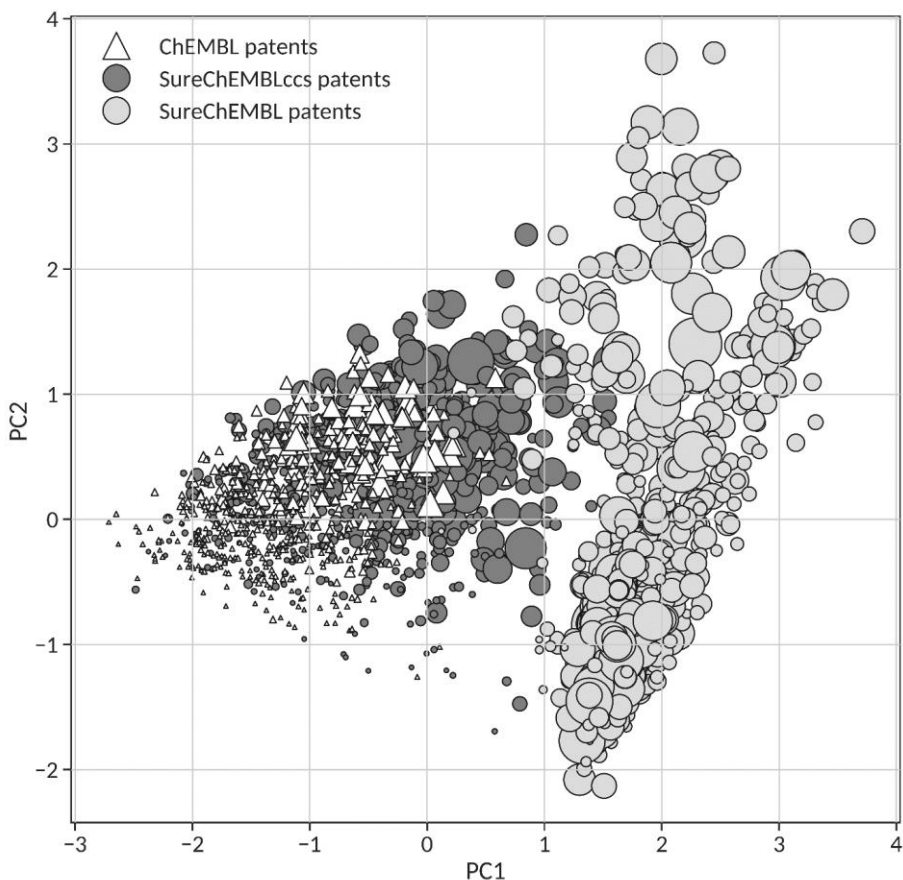


Figure 6. Principal Component Analysis of 750 patents based on the corresponding molecules available in ChEMBL (white triangles), SureChEMBLccs (dark grey circles) and SureChEMBL (light grey circles). PC1 and PC2 describe 75% and 23% of the total variance, respectively.

Figure 6 shows the projection of the 750 common patents between SureChEMBL (light grey circles), SureChEMBLccs (dark grey circles) and ChEMBL (white triangles) on the first two principal components that combined accumulate 98% of the variance (PC1 75% and PC2 23%). The loadings of the extent, mode and density values in PC1 (-0.65, -0.59 and -0.48, respectively) reveal a major contribution of extent and mode values in

Results

the first principal component, whereas the corresponding loadings in PC2 (0.14, 0.53 and -0.84, respectively) denote a major contribution of density values. The fact that PC1 describes already 75% of the variance allows the ordering of patents from all sources according to their intrinsic congenericity from left to right in the PC1 axis. Indeed, as can be observed, there is clear separation between the highly congeneric sets of claimed molecules assigned to all 750 patents in ChEMBL and SureChEMBLccs, on the left ($PC1 < 1$), and all molecules originally extracted from those patents and deposited in SureChEMBL, on the right ($PC1 > 1$). The strong presence of starting materials and intermediate products in SureChEMBL is certainly responsible for the low degree of congenericity associated with the full set of patent compounds in SureChEMBL. As clearly visible in Figure 6, this situation was corrected in SureChEMBLccs through the automatic identification of those molecules in the patent sharing a core chemical structure,²⁶ resulting in sets of patent compounds with significantly higher congenericities, comparable to those observed for the curated sets contained in ChEMBL. In this respect, it ought to be stressed that the size values for the 750 patents in ChEMBL range from 2 to 787 molecules per patent with a median value of 23, significantly smaller than the median size of 111 molecules in SureChEMBLccs for those same 750 patents. In fact, the number of claimed compounds per patent in SureChEMBLccs is on average 7.5 times larger than in ChEMBL. Therefore, the fact that SureChEMBLccs overlap well with ChEMBL for 750 patents (Figure 6) provides confidence for the high degree of congenericity of the chemical series for all 188,795 pharmacological patents available in SureChEMBLccs.²⁶

Given the optimal split obtained between patents in ChEMBL and SureChEMBLccs, on one side, and in SureChEMBL, on the other side, a

congenericity score (CScore) was defined as the geometric mean of the three similarity distribution descriptors used in the PCA. The distribution of CScores from the molecules available in the three patent sources for the set of 750 patents is presented in Figure 7. As can be observed, all patents in SureChEMBL (except three) have CScores below 0.4. In contrast, all patents in ChEMBL and 737 patents (98%) in SureChEMBLccs obtained CScores above 0.4. Therefore, a CScore threshold of 0.4 is recommended to assume a minimum degree of congenericity within patent molecules.

To illustrate the difference between patents containing a highly congeneric set of compounds and patents exemplified with more diverse chemical structures, two patent examples from SureChEMBLccs having CScores above and below 0.4 are included in Figure 7. Patent US-8796310 refers to the invention of amino-pyridine-containing compounds as spleen tyrosine kinase (SYK) inhibitors. The centroid of the patent molecules in SureChEMBLccs is compound SCHEMBL14840516 and its structure matches perfectly the Markush structure of the patent claim. The extent, mode and density values of its centroid similarity distribution are 0.72, 0.87 and 0.81, respectively. The resulting CScore of 0.80 reflects that molecules exemplified in this patent do not deviate much from the Markush structure. Conversely, patent US-9085555 has a CScore of 0.39 in SureChEMBLccs, right below the recommended CScore threshold of 0.4. The patent claims a set of compounds around a Markush structure that allows a wide diversity of ring sizes and composition, linkers and functional groups. The centroid of the filtered patent molecules in SureChEMBLccs is compound SCHEMBL12480885 and the similarity distribution constructed around it returned extent, mode and density values of 0.25, 0.43 and 0.56, respectively. Based on these results, CScore values offer a good simple

Results

metric to assess the congenericity of claimed compounds in patent applications.

Finally, it ought to be stressed that, very much in agreement with the results presented above (Figures 1-3), a strong correlation ($r^2 = + 0.93$) was found between the CScores calculated from the extent, mode and density values obtained from centroid similarity distributions and all pairwise similarity distributions. Therefore, even though this work focused on the use of centroid similarity distributions to perform a congenericity analysis of molecular sets (Figures 4-7), comparable results would be obtained by using all pairwise similarity distributions instead.

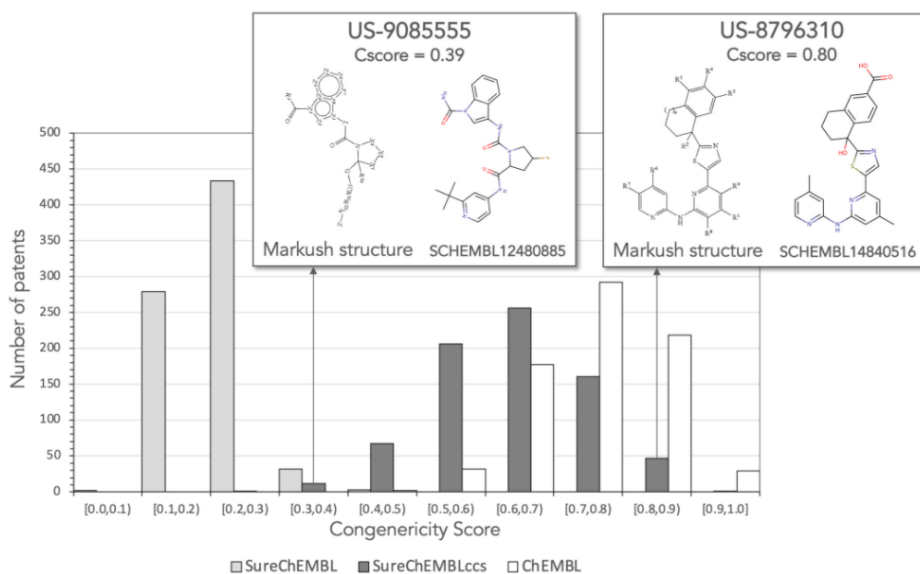


Figure 7. Congenericity Score (CScore) distributions for the same 750 patents in SureChEMBL, SureChEMBLccs and ChEMBL.

Conclusions

Pharmacological patent applications aim at protecting the invention of chemical series of compounds acting on the same mechanism-of-action target(s) or having similar cellular phenotype(s). Therefore, by definition, the sets of molecules claimed in patents can be chemically defined by a Markush structure with only small functional variations allowed or by a looser definition of a chemical series to cover the widest possible portion within that chemical space. Both patent strategies have their strengths and limitations and they could be balanced if a quantitative means to assess the degree of chemical compactness of all molecules contained in the patent would be available.

To this aim, a method was designed to calculate the degree of congenericity of claimed compounds in patent applications. The approach was applied and validated on a set of 750 patents from SureChEMBL for which a filtered set of molecules sharing a core chemical structure was available in SureChEMBLccs and a carefully curated set of at least two molecules was present also in ChEMBL. Patents were described by the similarity distribution around a reference compound and quantitatively characterized by its extent, mode and density values.

A principal component analysis (PCA) using the three similarity distribution descriptors successfully differentiated the patent molecular composition in each source, with filtered molecules in SureChEMBLccs showing overlapping congenericities with the manually curated sets in ChEMBL. A congenericity score (CScore), defined as the geometric mean of the extent, mode and density of similarity distributions, allowed for ranking patents according to the chemical compactness of their claimed molecules. Patent descriptors, CScores, and PC coordinates are provided as

Results

Supplementary Material to facilitate mapping future patents onto the PC space defined by the set of 750 curated patents. The current approach can be useful to describe the chemical space coverage of claimed compounds in pharmacological patent applications. More research in this direction is underway in our group.

Author Contributions: MJF carried out the work under the supervision of JM. The manuscript was written with contributions from both authors. All authors have given approval to the final version of the manuscript.

Funding: This work was partly supported by a project from the Spanish Ministerio de Ciencia e Innovación (PID2020-112539RB-I00).

Data availability statement: The filtered subset of molecules claimed by pharmacological US patents in SureChEMBL (SureChEMBLccs) is available for download at <ftp://ftp.ebi.ac.uk/pub/databases/chembl/SureChEMBLccs>. Patent descriptors (size, extent, mode and density), CScores and Principal Component coordinates (PC1 and PC2) of the 750 patents common in ChEMBL, SureChEMBLccs and SureChEMBL are provided as Supplementary Material.

Conflicts of interest: The authors declare no conflicts of interest.

References

1. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
2. Cases, M. *et al.* Chemical and biological profiling of an annotated compound library to the nuclear receptor family. *Curr. Top. Med. Chem.* **2005**, *5*, 763–772.
3. Schuffenhauer, A. *et al.* The Scaffold Tree – Visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* **2007**, *47*, 47–58.
4. Dalke, A.; Hert, J. & Kramer, C. Mmpdb: An open-source matched molecular pair platform for large multiproperty data sets. *J. Chem. Inf. Model.* **2018**, *58*, 902–910.
5. Bandyopadhyay, D. *et al.* Scaffold-based analytics: Enabling hit-to-lead decisions by visualizing chemical series linked across large datasets. *J. Chem. Inf. Model.* **2019**, *59*, 4880–4892.
6. Zhang, B.; Hu, Y. & Bajorath, J. AnalogExplorer: A new method for graphical analysis of analog series and associated structure-activity relationship information. *J. Med. Chem.* **2014**, *57*, 9184–9194.
7. Maggiora, G. M. On outliers and activity cliffs – Why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.
8. Hu, X.; *et al.* MMP-cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. *J. Chem. Inf. Model.* **2012**, *52*, 1138–1145.
9. Stumpfe, D.; Hu, H. & Bajorath, J. Advances in exploring activity cliffs. *J. Comput.-Aided Mol. Design* **2020**, *34*, 929–942.
10. Müller, G. Medicinal chemistry of target family-directed masterkeys. *Drug Discov. Today* **2003**, *8*, 681–691.

Results

11. Horton, D. A.; Bourne, G. T. & Smythe, M. L. The combinatorial synthesis of bicyclic privileged structures or privileged substructures. *Chem. Rev.* **2003**, *103*, 893–930.
12. Welsch, M. E.; Snyder, S. A. & Stockwell, B. R. Privileged scaffolds for library design and drug discovery. *Curr. Opin. Chem. Biol.* **2010**, *14*, 347–361.
13. Kim, J.; Kim, H. & Park, S. B. Privileged structures: efficient chemical “navigators” toward unexplored biologically relevant chemical spaces. *J. Am. Chem. Soc.* **2014**, *136*, 14629–14638.
14. Zhao, H. & Dietrich, J. Privileged scaffolds in lead generation. *Exp. Opin. Drug Discov.* **2015**, *10*, 781–790.
15. Davison, E. K. & Brimble, M. A. Natural product derived privileged scaffolds in drug discovery. *Curr. Opin. Chem. Biol.* **2019**, *52*, 1–8.
16. Bredel, M. & Jacoby, E. Chemogenomics: An emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **2004**, *5*, 262–275.
17. Mestres, J. Computational chemogenomics approaches to systematic knowledge-based drug discovery. *Curr. Opin. Drug Discov. Devel.* **2004**, *7*, 304–313.
18. Bajorath, J. Computational approaches in chemogenomics and chemical biology: current and future impact on drug discovery. *Expert Opin. Drug Discov.* **2008**, *3*, 1371–1376.
19. Stumpfe, D.; Dimova, D. & Bajorath, J. Computational method for systematic identification of analog series and key compounds representing series and their biological activity profiles. *J. Med. Chem.* **2016**, *59*, 7667–7676.
20. Naveja, J. J. *et al.* Systematic extraction of analogue series from large compound collections using a new computational compound-core relationship method. *ACS Omega* **2019**, *4*, 1027–1032.

21. Kruger, F.; Fechner, N. & Stiefl, N. Automated identification of chemical series: Classifying like a medicinal chemist. *J. Chem. Inf. Model.* **2020**, *60*, 2888–2902.
22. Heifets, A. & Jurisica, I. SCRIPDB: a Portal for Easy Access to Syntheses, Chemicals and Reactions in Patents. *Nucleic Acids Res.* **2012**, *40*, D428–D433.
23. Papadatos, G. *et al.* SureChEMBL: a Large-Scale, Chemically Annotated Patent Document Database. *Nucleic Acids Res.* **2016**, *44*, D1220–D1228.
24. Kunimoto, R. & Bajorath, J. Exploring Sets of Molecules from Patents and Relationships to other Active Compounds in Chemical Space Networks. *J. Comput. Aided Mol. Des.* **2017**, *31*, 779–788.
25. Akhondi, S. A. *et al.* Automatic Identification of Relevant Chemical Compounds from Patents. *A. Database (Oxford)* **2019**, baz001.
26. Falaguera, M. J. & Mestres, J. Identification of the core chemical structure in SureChEMBL patents. *J. Chem. Inf. Model.* **2021**, *61*, 2241–2247.
27. SureChEMBLccs **2021**.
<https://ftp.ebi.ac.uk/pub/databases/chembl/SureChEMBLccs> (last accessed on July 9th, 2021).
28. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*, D930–D940.
29. Landrum, G. A. RDKit: Open-source cheminformatics software, version 2017.09.1; <http://www.rdkit.org> (last accessed on December 13th, 2021).
30. Gregori-Puigjané, E. & Mestres, J. SHED: Shannon Entropy Descriptors from Topological Feature Distributions. *J. Chem. Inf. Model.* **2006**, *46*, 1615–1622.

Results

Supplementary Material

Sample of original Table S1. Patent descriptors (size, extent, mode and density), CScores and Principal Component coordinates (PC1 and PC2) for the 750 patents common in the three patent sources used (SureChEMBL, SureChEMBLccs and ChEMBL).

patent	dataset	size	extent	mode	density	cscore	PC1	PC2
US-20070135499	SureChEMBL	264	0.05	0.19	0.78	0.19	1.774	-0.902
US-20070135499	SureChEMBLccs	126	0.42	0.56	0.75	0.56	0.107	0.3
US-20070135499	ChEMBL	94	0.51	0.59	0.87	0.64	-0.658	-0.419
US-20070208166	SureChEMBL	346	0.04	0.28	0.54	0.18	2.531	0.949
US-20070208166	SureChEMBLccs	156	0.31	0.73	0.63	0.52	0.434	1.441
US-20070208166	ChEMBL	86	0.4	0.73	0.75	0.6	-0.258	0.658
US-20080096907	SureChEMBL	122	0.07	0.72	0.61	0.31	1.114	1.433
US-20080096907	SureChEMBLccs	43	0.69	0.77	0.86	0.77	-1.487	0.134
US-20080096907	ChEMBL	29	0.69	0.74	0.87	0.76	-1.454	-0.0
US-20080207655	SureChEMBLccs	24	0.7	0.71	0.86	0.75	-1.365	0.009
US-20080207655	ChEMBL	18	0.75	0.76	0.9	0.8	-1.765	-0.133
US-20080207655	SureChEMBL	120	0.04	0.14	0.69	0.16	2.276	-0.392
US-20080249081	ChEMBL	2	0.79	0.79	0.99	0.85	-2.291	-0.67
US-20080249081	SureChEMBL	369	0.15	0.28	0.76	0.32	1.395	-0.516
US-20080249081	SureChEMBLccs	67	0.64	0.72	0.85	0.73	-1.206	0.069
US-20100113462	ChEMBL	77	0.56	0.68	0.75	0.66	-0.52	0.633
US-20100113462	SureChEMBL	772	0.07	0.15	0.89	0.21	1.386	-1.739
US-20100113462	SureChEMBLccs	456	0.41	0.56	0.7	0.54	0.329	0.641
US-20100130505	ChEMBL	17	0.33	0.42	0.89	0.5	0.107	-1.019
US-20100130505	SureChEMBLccs	76	0.32	0.45	0.75	0.48	0.614	0.01

III.3 Illuminating the chemical space of untargeted proteins

Falaguera, M. J. & Mestres, J. Illuminating the chemical space of untargeted proteins. To be submitted.

Similar to the methodology introduced in Chapter III.1 to identify a core chemical structure representing the claim of pharmacological patents, in this article, a new approach is presented to identify those core scaffolds best representing the bioactive chemical series enriched within families of phylogenetically-related proteins. The obtained core scaffolds are then used to shed light on the chemical space of yet untargeted proteins included in the families.

Abstract

According to the Illuminating the Druggable Genome (IDG) initiative, 90% of the proteins encoded by the human genome still lack an identified biologically active ligand. Under this scenario, there is an urgent need for new approaches to chemically address these yet untargeted proteins. It is widely recognised that the best starting point for generating novel small molecules for proteins is to exploit the expected polypharmacology of known active ligands across phylogenetically related proteins following the paradigm that similar proteins are likely to interact with similar ligands. Here we introduce a computational strategy to identify core scaffolds that, when chemically expanded, are highly probable of containing active small molecules for untargeted proteins. The protocol was first tested on a set of 250 currently targeted proteins for which the year before their first reported active ligand there were at least two protein family members with known active ligands. A core scaffold contained in active ligands that were identified in the following years was correctly anticipated for 80 of those targeted proteins, a lower-bound performance estimate when considering data incompleteness. When applied to a set of 128 untargeted proteins, the identification of privileged core scaffolds present in known bioactive ligands of protein-family siblings allowed for extracting a priority list of commercially available small molecules. Assuming a minimum success rate of 32%, the chemical library selections should be able to deliver active ligands for at least 41 currently untargeted proteins.

Introduction

The Illuminating the Druggable Genome (IDG) initiative was launched in 2014 by the US National Institutes of Health (NIH) as an effort to quantify the amount of biomedical and pharmacological data available for human proteins and to ultimately increase our knowledge of the understudied human proteome.¹ Based on the type and amount of data available, proteins were assigned a target development level (TDL). Under the TDL scheme, it was found that only 10% of the proteins are mechanism-of-action targets of an approved drug (T_{clin}) or have at least one bioactive ligand deposited in public sources (T_{chem}). The remaining 90% of the human proteome is composed by chemically neglected proteins that nonetheless have well established implications in biological processes (T_{bio}) or their primary sequences is all what is currently known (T_{dark}).² Accordingly, illuminating the chemical space (ICS) of untargeted proteins remains a major challenge for the chemical biology and drug discovery communities and new approaches to accelerate current trends in protein deorphanisation and chemicalisation are needed.³⁻⁵

The TDL progression of proteins from $T_{\text{dark}}/T_{\text{bio}}$ to $T_{\text{chem}}/T_{\text{clin}}$ levels is not an easy endeavour and requires ample coordinated efforts. Some of the reasons hindering the identification of bioactive ligands for understudied proteins include the difficulty of identifying ligand binding sites in ligand-independent orphan targets,^{6,7} the absence of optimal assays to detect the activation of receptors with atypical coupling,^{8,9} and the absence of protein family members with already known bioactive ligands.⁵ All these aspects, added to increased high-throughput screening costs and more stringent safety regulations, justify that research in this field prefers investing on generating novel chemical series for already $T_{\text{chem}}/T_{\text{clin}}$ targeted proteins

Results

over initiating high risk projects on $T_{\text{dark}}/T_{\text{bio}}$ untargeted proteins.^{3,5} The development of computational strategies to define the chemical space likely to contain bioactive small molecules for a large number of untargeted proteins could have a major impact in promoting $T_{\text{dark}}/T_{\text{bio}}$ proteins to $T_{\text{chem}}/T_{\text{clin}}$ levels.

One of the most successful and frequently applied strategies to unveil the ligand space of $T_{\text{bio}}/T_{\text{dark}}$ proteins consists of using the chemical space of their phylogenetically-related proteins following the paradigm that similar proteins are likely to bind similar ligands.¹⁰⁻¹² Despite being a standard in drug discovery processes, especially in those within a polypharmacological setting,^{13,14} no systematic quantification of the effectiveness of this approach has been carried out to our knowledge. Neither a generalized protocol to be applied across different proteins families has been proposed beyond ligand similarity searches.^{15,16} With the aim of helping fill this gap, here, we introduce computational strategy to help ‘Illuminating the Chemical Space’ (ICS) of yet untargeted proteins based on the identification of privileged core scaffolds that best represent the chemical series enriched in bioactive compounds within the protein family of the untargeted protein. Applying a leave-one-out procedure to already $T_{\text{chem}}/T_{\text{clin}}$ targets, the protocol is evaluated on its ability to recover bioactive molecules containing those privileged core scaffolds. Then, the protocol is applied to true $T_{\text{dark}}/T_{\text{bio}}$ proteins that have phylogenetically close $T_{\text{chem}}/T_{\text{clin}}$ targets, and the privileged scaffolds obtained are then used to identify commercially-available compounds that are candidate bioactive ligands for the proteins. This final list of purchasable compounds for experimental testing is provided as Supplementary Material.

Methods

In this work we selected ChEMBL¹⁷ as a representative database of the ligand-target interactome currently known in the public domain, and ZINC¹⁸ as an open repository of purchasable and ready-to-test ligands. In its 28th release ChEMBL contains 4,440 human proteins and around 2 million small molecules. In its 20th release, ZINC contains over 90 million compounds commercially available for testing.

Target Development Level assignment. For these 4,440 human proteins, their associated ligand-protein bioactivities in ChEMBL were extracted. A protein having at least one small molecule (antibodies and biologics discarded) with a bioactivity value passing the family-specific thresholds specified by the IDG will be referred to as a targeted protein ($T_{\text{chem}}/T_{\text{clin}}$).¹ Moreover, given the purpose of our analysis, we added two extra requirements for a protein to be considered as targeted: (i) that the bioactivity data come from a direct single protein assay (ChEMBL confidence score equal to 9) and (ii) that the source documents from which the bioactivities were extracted are annotated with their publication year. All other proteins will be considered untargeted proteins ($T_{\text{dark}}/T_{\text{bio}}$).

Protein subfamily assignment. Hierarchical protein family classification codes were retrospectively generated for ChEMBL proteins. When the first levels of the code corresponded with a major protein family, they were abbreviated with the protein family prefix as follows: ‘1.6.1100.*’ code corresponds to protein kinases family and is abbreviated with the ‘KC.*’ prefix. According to this, proteins sharing the same full classification code (or until their penultimate level in the case of enzymes, nuclear receptors

Results

and cytochrome P450s) where classified in the same protein subfamily and defined as sibling proteins (Figure 1).

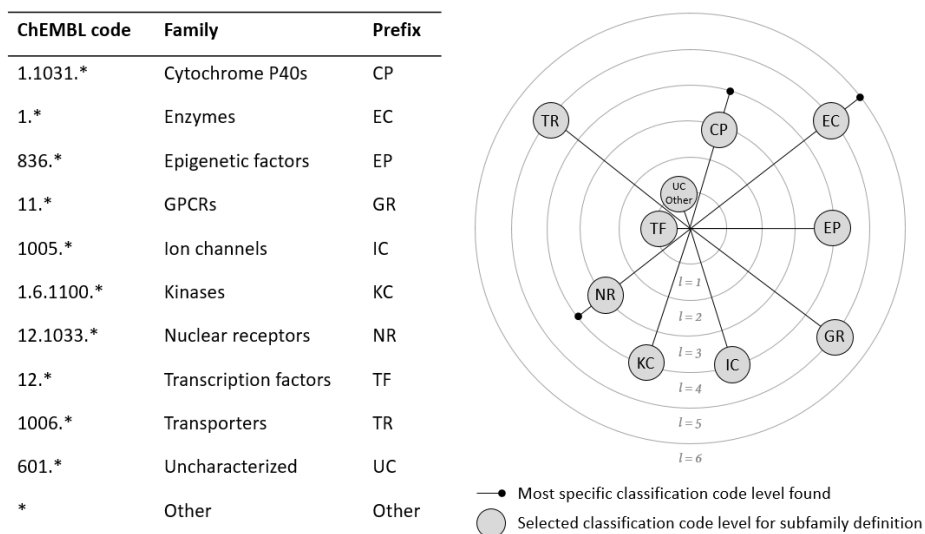


Figure 1. List of protein families (left) and protein subfamily assignment (right).

ICS protocol. The protocol to define the chemical space likely to be enriched in bioactive ligands for untargeted proteins consists of 7 steps as follows:

(1) *Protein subfamily assignment.* Assign the protein to a protein subfamily containing at least two $T_{\text{clin}}/T_{\text{chem}}$ targets to be used as reference.

(2) *Bioactive molecules collection.* For these $T_{\text{clin}}/T_{\text{chem}}$ sibling targets, collect their bioactive molecules available in ChEMBL fulfilling the conditions mentioned above.

(3) *Molecules to frameworks conversion.* Convert these siblings' bioactive molecules into their molecular frameworks, defined as the initial molecule

with all bonds converted into single ones and all atoms converted into carbons. In this way, molecular heterogeneity is reduced.

(4) *MCS of similar frameworks calculation.* Calculate the Maximum Common Substructures (MCSs) for those pairs of molecular frameworks coming from different siblings in the subfamily, that have a Dice similarity between their Morgan fingerprints with radius 2, equal or above 0.8. For MCS calculation, the RDKit¹⁹ function *rdFMCS.FindMCS* with the parameter *ringMatchesRingOnly* activated is used.

(5) *Core frameworks extraction.* Rank MCSs by the amount of sibling targets that have a bioactive molecule covered by it, and by the total amount of siblings' molecules covered. Then, collapse low-ranked MCSs into high-ranked ones when the second ones are substructures of the first ones. Select the top 5 ones to constitute the collection of the named 'core frameworks'.

(6) *Core scaffolds extraction.* Decorate the core frameworks with the original bond-types and atom-types of the substructures of the original molecules covered by them, to recover the molecular specificity reduced at step (3). The set of decorated frameworks will constitute the named 'core scaffolds'.

(7) *Candidate compounds screening.* Use the core scaffolds to screen a repository of purchasable compounds, such as ZINC, in search of candidate compounds to be tested against the untargeted protein. In this case, we used the has-substructure function from the RDKit PostgreSQL cartridge with the *adjustDegree* parameter deactivated and the *adjustRingCount* parameter activated.

Results and discussion

Target coverage. Following the criteria specified above, we found that from the 4,440 protein targets in ChEMBL, 4,224 (95%) have annotated bioactivities to small molecules. From them, 3,042 (69%) have quantitative bioactivities with a pChEMBL value annotated and 2,673 (60%) have a pChEMBL value passing the family-specific threshold. From the resulting set, only 2,262 (51%) come from assays with confidence score equal to 9, and from them 1,929 (43%) come from a source document with the year of publication annotated. This final collection of 1,929 (43%) proteins constitute our $T_{\text{clin}}/T_{\text{chem}}$ dataset, while the remaining 2,511 (57%) proteins are classified as $T_{\text{bio}}/T_{\text{dark}}$ (Table 1). Analysing these datasets across protein families, we found that most of $T_{\text{bio}}/T_{\text{dark}}$ are classified as enzymes (768) or in other protein families beyond the major ones (1,257), including proteins with family classifications annotated as ‘Uncharacterized’.

Among the proteins collected, there were identified a total of 144 protein subfamilies with at least two $T_{\text{clin}}/T_{\text{chem}}$ sibling targets, containing a median of 2 to 5 siblings with the exception of the transcription factors family, where all the 40 siblings are grouped in the same subfamily (Figure 1). These 144 families group 31% (599) of the total amount of $T_{\text{clin}}/T_{\text{chem}}$ proteins and only 5% (128) of $T_{\text{bio}}/T_{\text{dark}}$ ones (Table 2, Figure 2). The remaining 69% and 95% were not assigned to any subfamily. In all the protein subfamilies analysed the fraction of targets assigned to a subfamily within $T_{\text{dark}}/T_{\text{bio}}$ proteins is much smaller than within $T_{\text{clin}}/T_{\text{chem}}$ ones indicating that the lack of information associated to untargeted proteins occurs not only at the chemical annotation level but also at the physiological function annotation level, manifested in the absence of specific protein classification annotations for them. This notwithstanding, there is still a set of 128 yet untargeted

human proteins that can be assigned to a protein subfamily of at least two $T_{\text{clin}}/T_{\text{chem}}$ sibling targets, to which our bioactive chemical space prediction protocol can be applied.

Table 1. Distribution of $T_{\text{clin}}/T_{\text{chem}}$ and $T_{\text{bio}}/T_{\text{dark}}$ targets in ChEMBL.

Protein family	Targets	$T_{\text{clin}}/T_{\text{chem}}$	$T_{\text{bio}}/T_{\text{dark}}$
Cytochrome P450s	34	25 (74%)	9 (26%)
Enzymes	1,559	791 (51%)	768 (49%)
Epigenetic factors	132	81 (61%)	51 (39%)
GPCRs	357	241 (68%)	116 (32%)
Ion channels	212	118 (56%)	94 (44%)
Kinases	421	321 (76%)	100 (24%)
Nuclear receptors	47	37 (79%)	10 (21%)
Transcription factors	40	12 (30%)	28 (70%)
Transporters	150	72 (48%)	78 (52%)
Other	1,488	231 (16%)	1,257 (84%)
Total	4,440	1,929 (43%)	2,511 (57%)

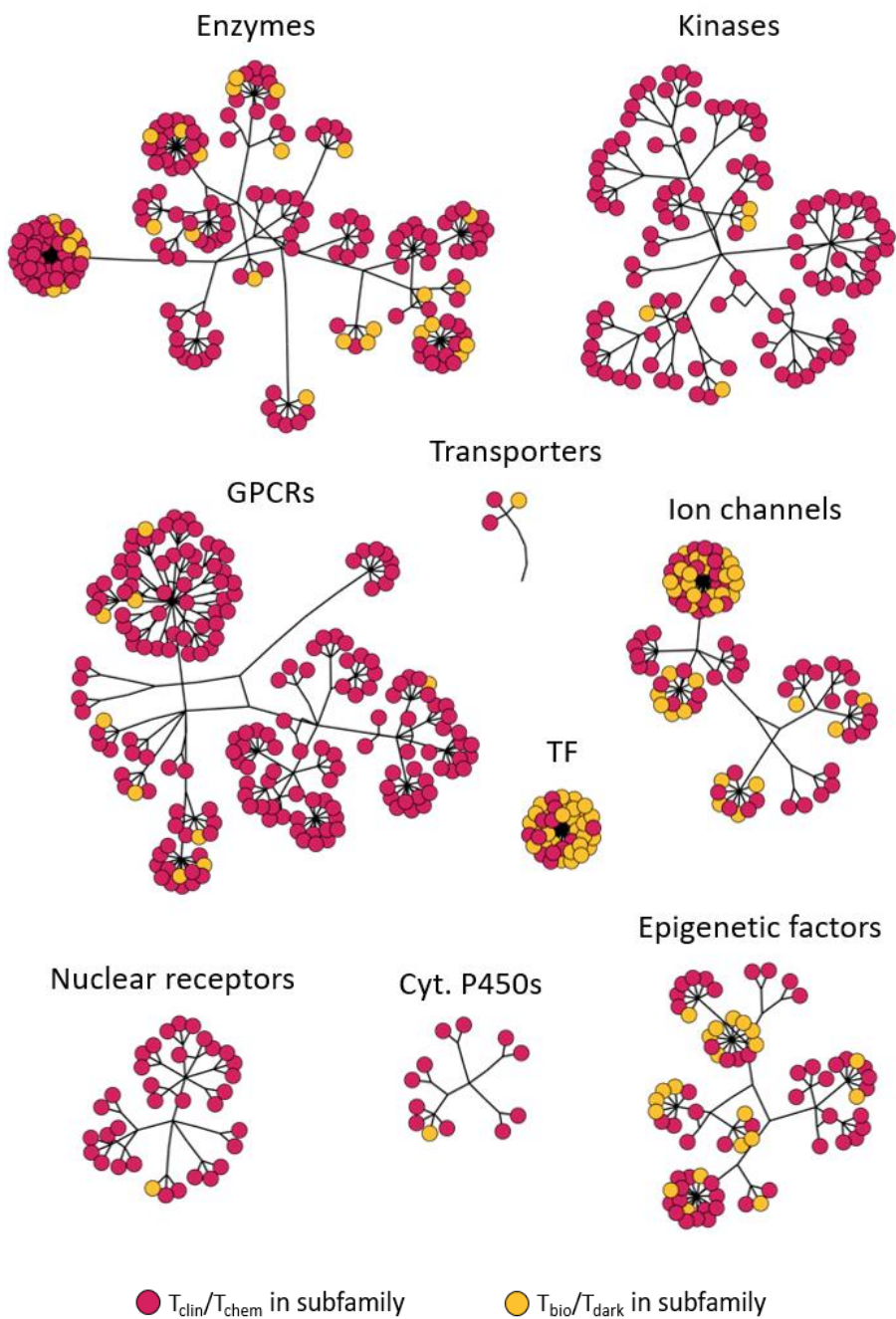


Figure 2. Distribution of 599 T_{clin}/T_{chem} and 128 T_{bio}/T_{dark} targets assigned to a protein subfamily. GPCRs, G-protein coupled receptors. TF, transcription factors.

Table 2. Distribution of $T_{\text{chem}}/T_{\text{clin}}$ and $T_{\text{dark}}/T_{\text{bio}}$ targets assigned to a protein subfamily.

Protein family	Subfamilies	Targets in subfamily	$T_{\text{clin}}/T_{\text{chem}}$ in subfamily	$T_{\text{bio}}/T_{\text{dark}}$ in subfamily
Cytochrome P450s	5	12 (35%)	11 (44%)	1 (11%)
Enzymes	23	184 (12%)	155 (20%)	29 (4%)
Epigenetic factors	13	72 (55%)	51 (63%)	21 (41%)
GPCRs	42	187 (52%)	178 (74%)	9 (8%)
Ion channels	10	92 (43%)	58 (49%)	34 (36%)
Kinases	37	103 (24%)	99 (31%)	4 (4%)
Nuclear receptors	12	34 (72%)	33 (89%)	1 (10%)
Transcription factors	1	40 (100%)	12 (100%)	28 (100%)
Transporters	1	3 (2%)	2 (3%)	1 (1%)
Other	0	0 (0%)	0 (0%)	0 (0%)
Total	144	727 (16%)	599 (31%)	128 (5%)

Results

Validation against $T_{\text{clin}}/T_{\text{chem}}$ proteins. To evaluate the efficacy of the protocol in identifying a collection of core scaffolds that represent the chemical series characteristic of the bioactive chemical space of a protein subfamily, and that are able to predict bioactive molecules for $T_{\text{bio}}/T_{\text{dark}}$ proteins, we performed a virtual de-targeting exercise. This consists of (1) selecting as a validation dataset current $T_{\text{clin}}/T_{\text{chem}}$ that were found within a protein subfamily of at least two $T_{\text{clin}}/T_{\text{chem}}$ sibling targets the year before the first bioactive molecule was deposited for them in ChEMBL, (2) applying the prediction protocol to the chemical space known for their sibling targets at that year, and (3) using the obtained siblings' core scaffolds to try to recover any of the bioactive molecules currently annotated to them. The results of this exercise are shown in Table 3. The validation dataset is composed by 250 $T_{\text{clin}}/T_{\text{chem}}$ targets (13% of initial ones), most of them classified as enzymes (98) or GPCRs (69). For 82% of them (205), a core scaffold was identified, indicating that for the remaining 18% (45) no inter-sibling pairs of molecular frameworks with a similarity value equal or above 0.8 to extract an MCS were found. For half of the targets processed, between 1 and 6 MCS frameworks collapsed were obtained, with few outlier cases with up to 283 core frameworks corresponding to large protein subfamilies with very diverse chemical series associated. These core frameworks were found to be very specific for the subfamilies, most of them occurring in only 1 subfamily with the exception of the staurosporin scaffold appearing in 5 protein kinases subfamilies, as expected.

When trying to recover currently known bioactive molecules for these targets we relaxed the bioactivity cutoff for all the protein families to 4.6 (the lowest cutoff found in Oprea *et al.* (2017)¹) to get the most of the coverage. Then, using the siblings' core scaffolds obtained, we found out that in 32% (80) of the cases an active molecule was recovered, in 1 (<1%)

case only inactive (bioactivity below 4.6) molecules were recovered and in the remaining 50% (124) neither active nor inactive molecules were recovered. The low percentage of targets with only inactive molecules recovered (<1%) compared to the 32% of targets with bioactive molecules recovered shows the efficacy of our protocol in identifying bioactive chemical series for the protein subfamily and, thus for the false $T_{\text{bio}}/T_{\text{dark}}$ protein. On the other hand, when looking closer to the 50% of the cases where the core scaffolds derived from the siblings' chemical space were not able to recover any bioactive molecule of the target analysed, we found that this lack of cross-targeting between targets that are known to be sequence- and structurally-similar is explained by the lack data completeness at the ligand-target interaction level, so widely discussed for years.^{20,21,22} Assuming the cross-pharmacology between sibling proteins in the same subfamily, researchers at the medicinal chemistry area tend to select single representative ones for the experimental testing in search for bioactive ligands and use them as sentinels²³ to project the results obtained to their siblings. Thus, it is very typical to find protein subfamilies with one target having many bioactivities associated compared to the few ones associated to the other siblings. Our results show that this completeness issue is particularly acute in protein families like enzymes, ion channels and transcription factors, with less than half of the targets analysed having at least one active molecule recovered. The design of efficient algorithms for the prediction of targets chemical space is dependent on ligand-target matrices completeness as shown here and in other publications,²⁴ to help mitigate this issue a complete list of the bioactive molecules associated to the $T_{\text{clin}}/T_{\text{chem}}$ targets considered for the validation, annotated to their pChEMBL value, their deposition year in ChEMBL and whether they are

Results

or not recovered by the siblings' core scaffolds extracted is provided in the Supplementary Table S1.

Table 3. Illuminating the Chemical Space protocol applied to 250 $T_{\text{clin}}/T_{\text{chem}}$ proteins one year before their first active molecule was deposited in ChEMBL. Percentages are calculated with respect to the targets in the first column.

Protein family	Targets	Targets with core scaffolds identified	Targets with molecules recovered	Targets with no molecules recovered
Cytochrome P450s	1	1 (100%)	1 (100%)	0 (0%)
Enzymes	98	87 (89%)	23 (23%)	63 (64%)
Epigenetic factors	19	17 (89%)	12 (63%)	5 (26%)
GPCRs	69	61 (88%)	32 (46%)	29 (42%)
Ion channels	34	23 (68%)	2 (6%)	21 (62%)
Kinases	16	10 (63%)	9 (56%)	1 (6%)
Nuclear receptors	3	1 (33%)	1 (33%)	0 (0%)
Transcription factors	10	5 (50%)	0 (0%)	5 (50%)
Transporters	0	0 (-)	0 (-)	0 (-)
Other	0	0 (-)	0 (-)	0 (-)
Total	250	205 (82%)	80 (32%)	124 (50%)

In Figure 3 some illustrative examples of the protocol validation across different protein families are shown. The first one corresponds to the coagulation factor IX/VIII (*F9*) classified in the serine protease S1A subfamily (EC.1028.1079.1136.38) and with the first active ligand deposited in ChEMBL in 2004. In 2003 (2004–1), there were 17 $T_{\text{clin}}/T_{\text{chem}}$ sibling targets, including the thrombin and coagulation factor VII (*F7*), the thrombin and coagulation factor X (*F10*) and the thrombin and trypsin (*F2*). The top core framework found for them at that time covered molecules annotated to 7 of these siblings and with only one of the two core scaffolds derived from this top1 core framework we are able to recover one of the bioactive molecules currently associated to *F9* in ChEMBL, named CHEMBL327715. When considering the final set of all core scaffolds derived from the top5 core frameworks, a total of 4 bioactive molecules currently associated to the false $T_{\text{bio}}/T_{\text{dark}}$ protein are recovered showing the efficacy of our protocol. The second example corresponds to the D(1A) dopamine receptor (*DRD1*) classified in the dopamine receptors subfamily (GR.1020.1088.1266.535) and with the first active ligand deposited in ChEMBL in 1996. In 1995 (1996–1), there were 3 $T_{\text{clin}}/T_{\text{chem}}$ sibling targets, the D(2) dopamine receptor (*DRD2*), the D(3) dopamine receptor (*DRD3*) and the D(4) dopamine receptor (*DRD4*). There were obtained 3 core frameworks for them at that year that, when decorated, derived in 13 core scaffolds. The top ranked core framework was able to recover one of the bioactive molecules currently associated to *DRD1*, named CHEMBL54, and a second molecule was also recovered when considering all the top3 core frameworks, named CHEMBL2158640. This second example shows the efficacy of the protocol also for small protein subfamilies. The third example corresponds to the voltage-gated potassium channel subunit Kv1.6 (*KCNA6*) classified in the voltage-gated potassium channels subfamily

Results

(IC.1019.667.118), abbreviated as VG K, and with the first active ligand deposited in ChEMBL in 2017. In 2016 (2017–1), there were 17 $T_{\text{clin}}/T_{\text{chem}}$ sibling targets, including the voltage-gated potassium channel subunit Kv4.3 (*KCND3*), the voltage-gated potassium channel subunit Kv7.2 (*KCNQ2*) and the voltage-gated potassium channel subunit Kv1.4 (*KCNA4*). The top ranked core framework obtained for the siblings at 2016 derives into 6 core scaffolds with one of them (the one shown in fig. 3) recovering one of the bioactive molecules currently associated to *KCNA6*, named CHEMBL444449. The fourth and last example corresponds to the disintegrin metalloproteinase domain-containing protein 33 (*ADAM33*) classified in the metalloprotease M12B subfamily (EC.1028.1081.1131.84), abbreviated as M12B. The first active ligand for this target was deposited in ChEMBL in 2014, and in 2013 (2014–1) there were 5 $T_{\text{clin}}/T_{\text{chem}}$ sibling targets, including the disintegrin metalloproteinase domain-containing protein 10 (*ADAM10*), the disintegrin metalloproteinase domain-containing protein 17 (*ADAM17*) and the A disintegrin and metalloproteinase with thrombospondin motifs 5 (*ADAMTS5*). The top core framework identified for the sibling targets covered molecules annotated to 4 of these siblings at 2013 year and with only one of the two core scaffolds derived from this framework, one of the bioactive molecules currently associated to *ADAM33* in ChEMBL is recovered, named CHEMBL3643916. When considering the final set of all core scaffolds derived from the top5 core frameworks, a total of 9 bioactive molecules currently associated to the false $T_{\text{bio}}/T_{\text{dark}}$ protein are recovered showing the efficacy of our protocol.

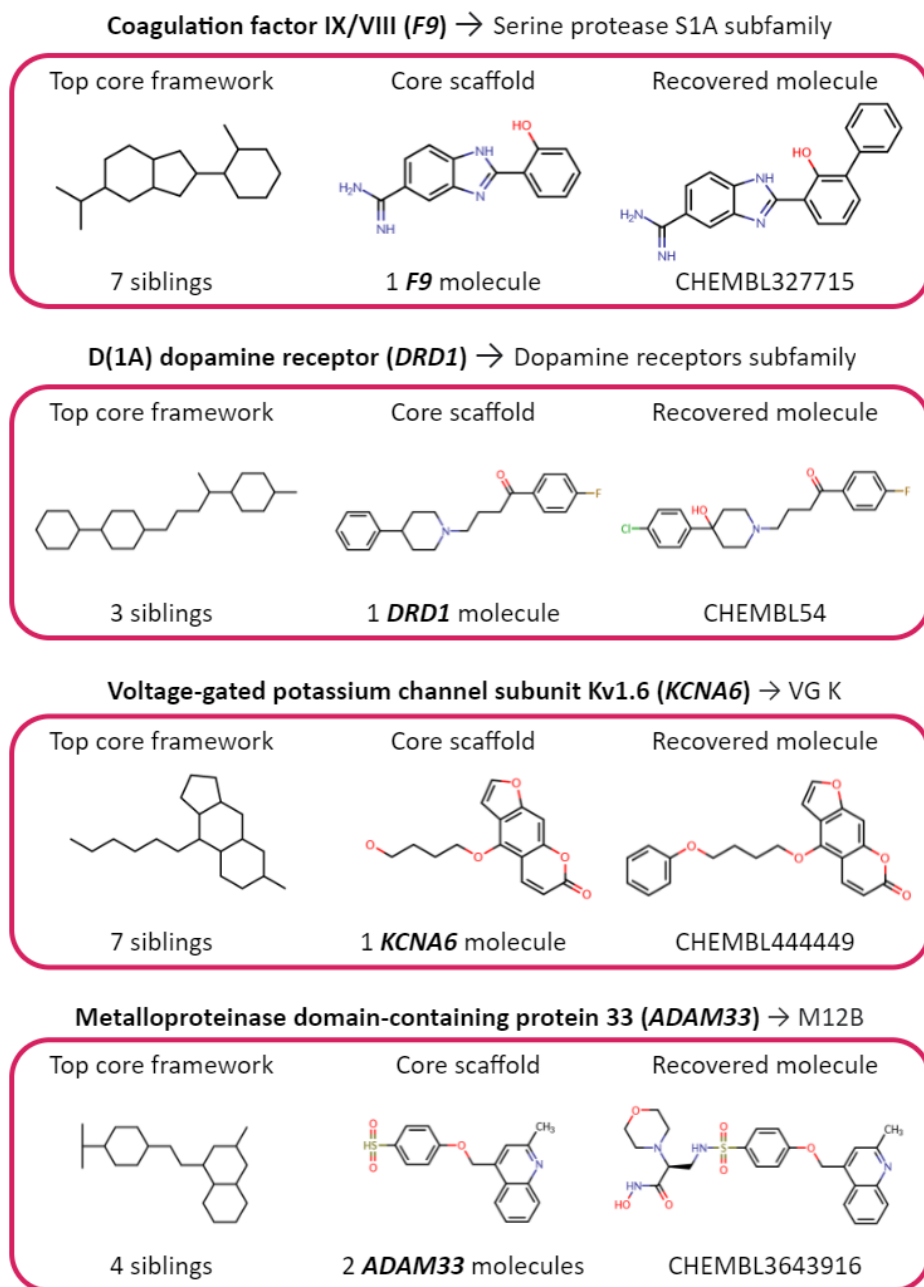


Figure 3. Illustrative cases of $T_{\text{clin}}/T_{\text{chem}}$ proteins in the validation dataset.

Results

Application to $T_{\text{bio}}/T_{\text{dark}}$ proteins. Having validated the performance of the prediction protocol on the 250 $T_{\text{clin}}/T_{\text{chem}}$ targets, the next step was to apply it to the set of 128 current $T_{\text{bio}}/T_{\text{dark}}$ proteins that can be assigned to a protein subfamily with at least two $T_{\text{clin}}/T_{\text{chem}}$ sibling targets. The results obtained are collected in Table 4. For 91% of them (116), a core scaffold was identified, above the 82% obtained for the validation dataset. The median amount of core frameworks per target increases with respect to the median amount obtained for the validation dataset to up to 46 indicating that the chemical space associated to protein subfamilies that contain $T_{\text{bio}}/T_{\text{dark}}$ proteins is more heterogeneous and cannot be collapsed into few MCSs frameworks. However, they are found again to be very specific for the subfamilies, all of them occurring in only 1 or two subfamilies.

The core scaffolds identified for the 116 $T_{\text{bio}}/T_{\text{dark}}$ targets sum up to 1,439 unique structures that come from 150 unique core frameworks. This core frameworks have a median amount of 5 core scaffolds each one. For the reader to have a feeling of the structures obtained, in Figure 4 we show a sample of the core scaffolds resulting from the selection of one representative scaffold of the top1 ranked core frameworks extracted for each of the unique subfamilies containing these 116 $T_{\text{bio}}/T_{\text{dark}}$ protein.

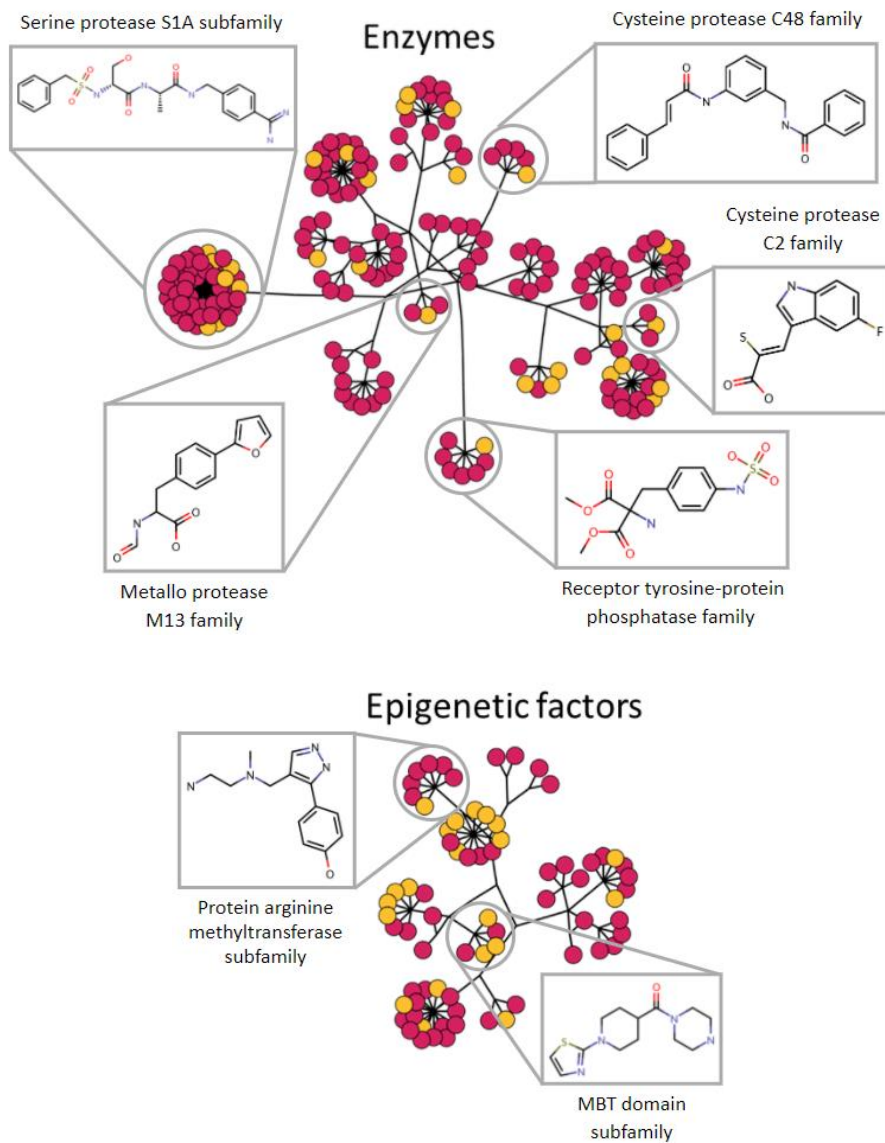


Figure 4. Sample of core privileged scaffolds obtained for a selection of the 128 untargeted protein in enzymes, ion channels, epigenetic factors and GPCRs.

Results

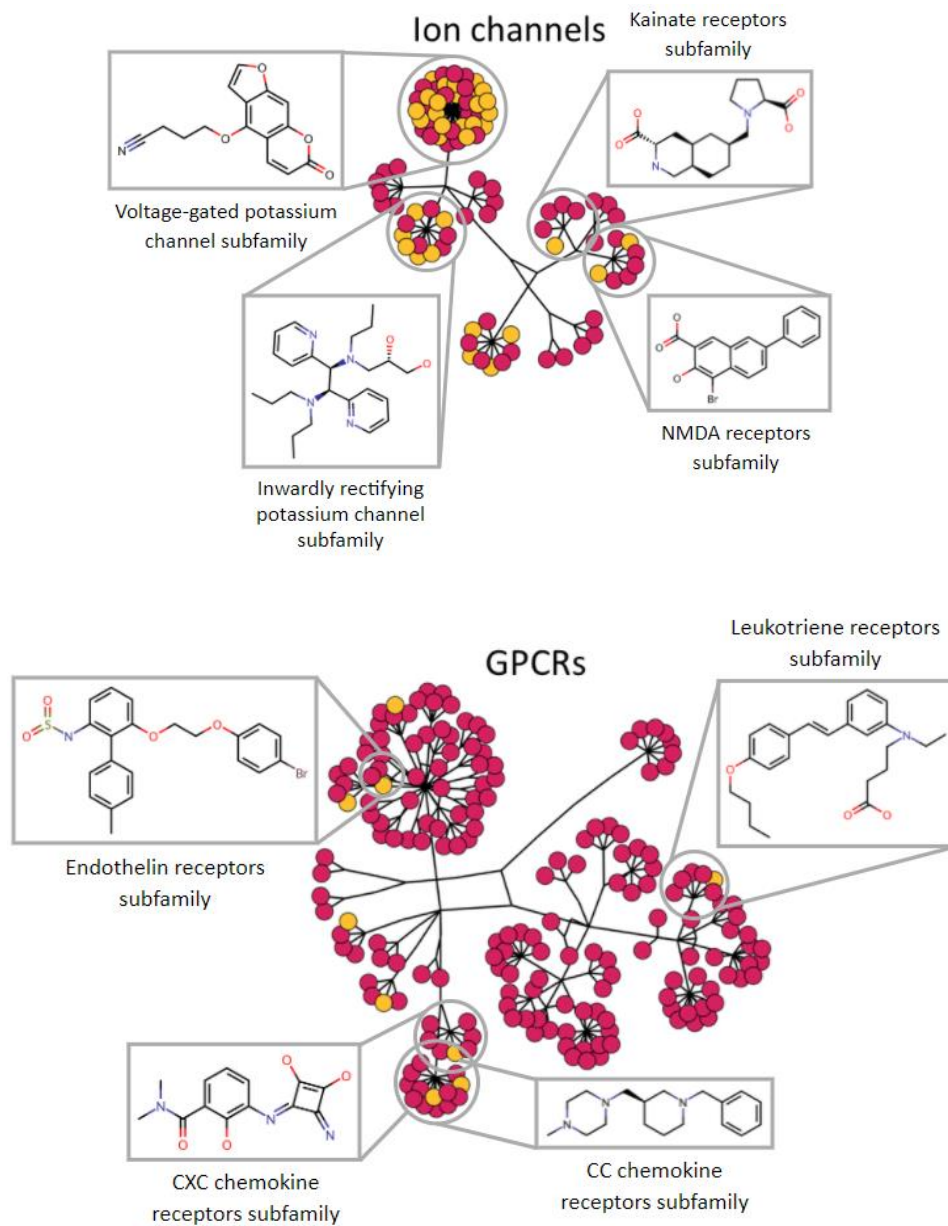


Figure 4. (continued)

Contrary to the results obtained when applying the molecular recovery step (step 7) to the validation dataset, which resulted in only 32% of the targets analysed having a bioactive molecule recovered by the siblings' core scaffolds, in this case at least one molecule in ZINC database could be identified with the core scaffolds extracted in up to 77% (98) of the $T_{\text{bio}}/T_{\text{dark}}$ targets processed. The ZINC molecules recovered sum a total of 494,274 (Table 4) unique compounds in-stock ready to be purchased and tested against the yet untargeted proteins. The probability of finding a bioactive molecule for the target among them equals $32\pm 50\%$, as we concluded from the validation analysis. The complete list of ZINC compounds recovered is provided in the Supplementary Table S2 mapped to their corresponding $T_{\text{bio}}/T_{\text{dark}}$ targets. For those targets with no ZINC compounds found but with a core scaffold identified, the collection of core scaffolds obtained are also provided as an starting point for *de novo* candidate drug design.

As a closing of this section, we selected an exemplary $T_{\text{bio}}/T_{\text{dark}}$ target for a detailed and step-by-step description of how the protocol is applied and which results are obtained. The selected protein is the lethal(3)malignant brain tumor-like protein 4 (*L3MBTL4*) classified within the MBT domain subfamily (EP.837.858.861) (fig. 5). Together with the *L3MBTL4* protein, this subfamily contains two $T_{\text{clin}}/T_{\text{chem}}$ siblings, the Lethal(3)malignant brain tumor-like protein 3 (*L3MBTL3*), with the first deposited bioactivity in ChEMBL in 2013, and the Lethal(3)malignant brain tumor-like protein 1 (*L3MBTL1*), with the first deposited bioactivity in ChEMBL in 2017. When collecting ChEMBL bioactivities for the sibling targets we found that *L3MBTL3* has 55 active molecules annotated while *L3MBTL1* has 1,118, evidencing the completeness bias towards one target in the subfamily as the representative ones mentioned before. Among the molecules collected, a 3-

Results

ring PS is detected in both siblings' chemical space. This substructure is exemplified and labelled in orange in Figure 4 step (2) in the molecule CHEMBL2426373 associated to *L3MBTL3* with a pIC50 equal to 6.46, and in CHEMBL1348716 associated to *L3MBTL1* with a potency value equal to 8.7. Apart from this chemical series, many others are found associated to *L3MBTL1* which we exemplify with the CHEMBL1257003 molecule that has a pIC50 equal to 6.55. As it can be appreciated, the PS highlighted has a constant molecular framework but different atom and bond types decorating the ring. This justifies step (3) molecule-to-framework conversion in order to increase the MCS signal at step (4) and facilitate the detection of an inter-siblings enriched substructure. At MCS calculation step, CHEMBL1257003 is left behind and only CHEMBL2426373 and CHEMBL1348716 molecules are retained for the next steps. In this particular case only one framework is shown so it is selected as the single core framework at step (5) covering 2 out of the 2 siblings clustered in the MBT domain subfamily. Once the core framework is identified, it is decorated back to the initial substructures where it was found in the original molecules to give rise to the collection of final core scaffolds exemplified here with the two substructures at step (6). With them, ZINC database was screened and we discovered 61 molecules containing the core scaffolds as substructures, such as ZINC19715644 and ZINC12543865. These molecules are ready to be purchased for testing against *L3MBTL4*.

Table 4. Illuminating the Chemical Space protocol applied to 128 $T_{\text{bio}}/T_{\text{dark}}$ proteins. Percentages are calculated with respect to the targets in the first column.

Protein family	Targets	Targets with core scaffold found	Targets with ZINC molecules identified	Unique ZINC molecules identified
Cytochrome P450s	1	1 (100%)	0 (0%)	0
Enzymes	29	28 (97%)	27 (93%)	317,814
Epigenetic factors	21	13 (62%)	7 (33%)	164,819
GPCRs	9	9 (100%)	5 (56%)	3,070
Ion channels	34	34 (100%)	28 (82%)	101
Kinases	4	3 (75%)	3 (75%)	72
Nuclear receptors	1	0 (0%)	0 (0%)	0
Transcription factors	28	28 (100%)	28 (100%)	8,398
Transporters	1	0 (0%)	0 (0%)	0
Other	0	0 (-)	0 (-)	0
Total	128	116 (91%)	98 (77%)	494,274

Results

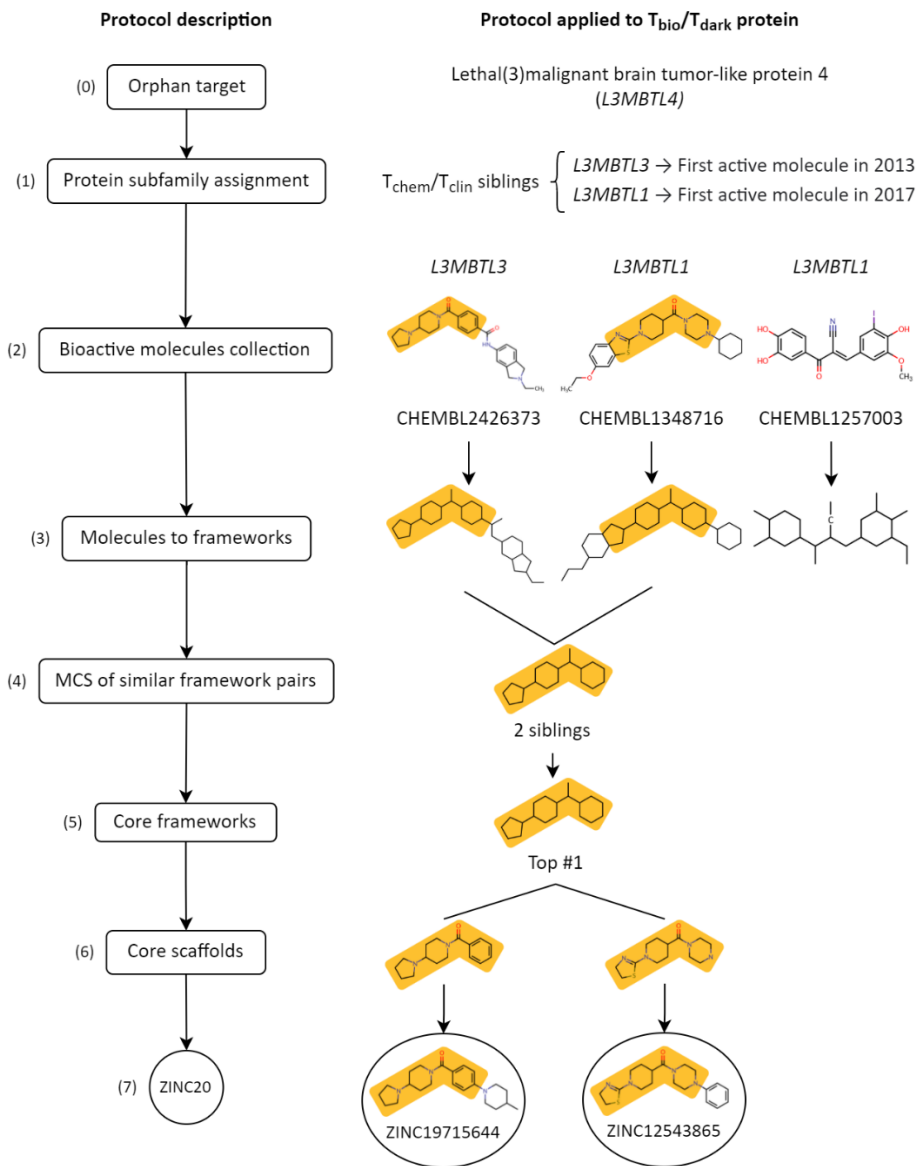


Figure 5. Illustrative example of prediction protocol applied to a $T_{\text{bio}}/T_{\text{dark}}$ protein.

Conclusions

It is widely established that if a small molecule binds to a given protein, it may bind also to other proteins related by sequence identity. Target phylogeny is thus a property worth exploiting in the quest for bioactive small molecules of untargeted proteins belonging to protein families with already targeted protein members. The connection between target phylogeny and ligand polypharmacology motivated more than a decade ago the implementation of chemogenomics strategies in drug discovery, aiming at organising research around target families as a means to maximise efficiency of chemistry and biology resources and to improve hit rates.¹⁰⁻¹²

In a sense, this work is an attempt to recover the lost chemogenomics spirit. By maximally exploiting the structural contents of the growing number of bioactive ligands contained in public sources, we have been able to define a computational strategy to identify privileged core scaffolds likely to be enriched with bioactive ligands across multiple members of a protein family. In those protein families with untargeted proteins, the information on family-wide privileged scaffolds is then used to extract commercially available small molecules containing them. It is expected that these focussed chemical sets will accelerate the discovery of bioactive ligands for many untargeted proteins. All molecular sets are made publicly available for the benefit of the entire chemical biology and drug discovery communities.

Acknowledgments

This work was partly supported by a project from the Spanish Ministerio de Ciencia e Innovación (PID2020-112539RB-I00)

Supplementary Material

Supplementary Table S1

File: chembl28_Tclin_Tchem.csv.gz

T_{clin}/T_{chem} | chembl_id | pChEMBL | year | recovered (True/False)

Supplementary Table S2

File: chembl_28_Tbio_Tdark.csv.gz

T_{bio}/T_{dark} | core_scf | zinc_id | SMILES

References

1. Oprea, T. *et al.* Unexplored therapeutic opportunities in the human genome. *Nat Rev Drug Discov.* **2018**, *17*, 317–332.
2. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*(D1), D480–D489.
3. Tunaru, S. Strategies for G-protein coupled receptor deorphanization. *Molecular Life.* **2017**, *1*(1), 71–79.
4. Levoe, A. & Jockers, R. Alternative drug discovery approaches for orphan GPCRs. *Drug discovery today* **2008**, *13*(1–2), 52–58.
5. Laschet, C.; Dupuis, N. & Hanson, J. The G protein-coupled receptors deorphanization landscape. *Biochem Pharmacol.* **2018**, *153*, 62–74.
6. Davenport, A. P. *et al.* International Union of Basic and Clinical Pharmacology. LXXXVIII. G protein-coupled receptor list: recommendations for new pairings with cognate ligands. *Pharmacol Rev.* **2013**, *65*(3), 967–986.

7. Levoye, A. *et al.* Do orphan G-protein-coupled receptors have ligand-independent functions? New insights from receptor heterodimers. *EMBO Rep.* **2006**, 7(11), 1094–1098.
8. Rajagopal, S. *et al.* Beta-arrestin- but not G protein-mediated signaling by the "decoy" receptor CXCR7. *Proc Natl Acad Sci U S A.* **2010**, 107(2), 628–632.
9. Okinaga, S *et al.* C5L2, a nonsignaling C5A binding protein. *Biochemistry.* **2003**, 42(31), 9406–9415.
10. Bredel, M. & Jacoby, E. Chemogenomics: an emerging strategy for rapid target and drug discovery. *Nat. Rev. Genet.* **2004**, 5(4), 262–275.
11. Mestres, J. Computational chemogenomics approaches to systematic knowledge-based drug discovery. *Curr. Opin. Drug Discov. Devel.* **2004**, 7(3), 304–313.
12. Klabunde, T. & Hessler, G. Drug design strategies for targeting G-protein-coupled receptors. *Chembiochem.* **2002**, 3(10), 928–944.
13. Müller, G. Medicinal chemistry of target family-directed masterkeys. *Drug Discov. Today* **2003**, 8, 681–691.
14. Meyers, J. *et al.* Privileged Structures and Polypharmacology within and between Protein Families. *ACS Med Chem Lett.* **2018**, 9(12), 1199–1204.
15. Schuffenhauer, A.; Floersheim, P.; Acklin, P. & Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J Chem Inf Comput Sci.* **2003**, 43(2), 391–405.
16. Wassermann, A. M.; Geppert, H. & Bajorath, J. Ligand prediction for orphan targets using support vector machines and various target-ligand kernels is dominated by nearest neighbor effects. *J Chem Inf Model.* **2009**, 49(10), 2155–2167.
17. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, 47(D1), D930–D940.

Results

18. Irwin, J. J. & Shoichet, B. K. ZINC--a free database of commercially available compounds for virtual screening. *J Chem Inf Model.* **2005**, *45*(1), 177–82.
19. Landrum, G. A. RDKit: Open-source cheminformatics software, version 2017.09.1; <http://www.rdkit.org> (last accessed on December 13th, 2021).
20. Nickerson, R. S. Confirmation bias: a ubiquitous phenomenon in many guises. *Rev. Gen. Psychol.* **1998**, *2*, 175–220.
21. Mestres, J.; Gregori-Puigjané, E.; Valverde, S. & Solé, R. V. Data completeness--the Achilles heel of drug-target networks. *Nat Biotechnol.* **2008**, *26*(9), 983–4.
22. Gregori-Puigjané, E. & Mestres, J. Coverage and bias in chemical library design. *Curr Opin Chem Biol.* **2008**, *12*(3), 359–365.
23. Bowes, J. *et al.* Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat Rev Drug Discov.* **2012**, *11*(12), 909–922.
24. Kanai, C. *et al.* Computational prediction of compound-protein interactions for orphan targets using CGBVS. *Molecules.* **2021**, *26*(17), 5131.

III.4 A map of the proteome targetable by dual-acting agents

Falaguera, M. J. & Mestres, J. A map of the proteome targetable by dual-acting agents. In preparation.

As a closing chapter of this section, here I present a new ontology to help map current poly-pharmacological opportunities for targeting the human proteome by dual-acting agents.

Introduction

In the last 20 years, polypharmacology¹ has emerged as a very promising therapeutic strategy for the treatment of complex diseases such as cancer and central nervous system (CNS) disorders, with demonstrated efficacy and safety improvements compared to magic bullet-based therapies and drug combinations.² The intricate framework of protein-protein interactions (PPI), protein-metabolite interactions, compensatory signalling routes and feedback mechanisms^{3,4} that characterize these kind of diseases ensures the robustness of the systems when one of its elements is perturbed⁵ and makes extremely difficult, if not impossible, for single-action therapeutics to overcome the system resistance. Although the standard treatments in this kind of settings have for years consisted in the administration of drug combinations ('drug cocktails'),⁶ safety issues derived from drug-drug interactions, negative synergistic effects and dosage selection that are associated to drug mixtures² is making multi-action drugs to be considered more and more the suitable alternative for therapy.

As the knowledge on the molecular mechanisms behind drugs action increases, it is becoming more evident that many drugs approved in the past as single-action ones, in fact owe their efficacy to their multi-action behaviour.⁷ This is prompting a shift in the perception of drug off-targets effects in the medicinal chemistry community. From being considered as prone to safety issues, undesirable and necessary to be removed, they are progressively been seen as promising opportunities for multi-target drugs' discovery.⁷

The simplest level of polypharmacology applied to drug discovery are dual-action drugs (DADs). These are compounds approved on purpose to

Results

act simultaneously on pairs of protein targets at similar efficacious dose.⁹ Some examples of DADs approved in the last years include bosutinib, a dual SRC/ABL kinase inhibitor for the treatment of chronic myeloid leukaemia;¹⁰ lapatinib, a dual EGFR/ERBB2 inhibitor for the treatment of breast cancer;¹¹ bupropion, a dual serotonin/dopamine receptors blocker used as antidepressant;¹² and clozapine, another dual serotonin/dopamine receptors blocker approved as an antipsychotic.¹³ And several other agents are at research and pre-clinical phases as clinical candidates for a variety of diseases.⁹

Although some attempts to globally map the ligand-target space of some therapeutically relevant protein families have been made in the last 20 years,^{7,14,15} no systematic assessment of dual-action, and by extent, multi-action drug opportunities across the whole targetable proteome has been carried out to our knowledge. Thus, starting from the lowest level, and with the possibility of extending it to higher ones, here we propose a novel whole ontology to help describe target *vs* target links space in the context of DADs possibilities and the derived space of dual-agent *vs* target *vs* target triads. The described ontology is applied to analyse current dual-pharmacology situations and opportunities for the targetable proteome defined by the Illuminating the Druggable Genome (IDG) initiative.¹⁶

The paper is organized as follows, first we define different categories to describe what we name as dual agents, according to their mechanism of action (MoA) association with the proteins targeted. Secondly, we define different categories to describe target pairs categories according to their target development level (TDL)¹⁶ proposed by the IDG. Finally, we use the defined concepts to build a map of the proteome targetable by dual-acting agents and discuss future opportunities for dual pharmacology.

Dual agents categories

We define a *dual agent* as a compound with biologically relevant activity (see Databases and Methods) for a pair of protein targets. They can be subclassified in three categories according to their maximum (pre)clinical phase and their mechanism of action association with the pair of proteins targeted, as follows:

- **Dual action drug (DAD)**: approved drug designed on purpose to combine two different pharmacological actions at similar efficacious dose.
- **Dual interacting drug (DID)**: approved drug not designed on purpose as a DAD, which has clinically-relevant activity for two protein targets. They can be subclassified in three subcategories according to the pair of targets covered:
 - **DID_{MOA}**, if both targets are defined as its MoA (putative DADs);
 - **DID_{MIX}**, if only one of the targets is defined as its MoA; and
 - **DID_{OFF}**, if none of the targets are defined as its MoA, but as off-targets.
- **Dual interacting ligand (DIL)**: non-approved ligand with clinically-relevant activity for two protein targets.

Target pairs categories

The TDL system designed by the IDG initiative classifies human proteins into four categories (named T_{clin} , T_{chem} , T_{bio} and T_{dark}) according to their known biomedical and pharmacological relevance. T_{clin} (clinic) targets refer to those having an approved drug annotated as its mechanism of action (MoA). T_{chem} (chemistry) proteins stand for those that lack a MoA-based link to approved drugs but that are known to bind to small molecules with high potency. Ligand-target interaction bioactivities for T_{clin} vs T_{chem} discrimination are extracted from ChEMBL¹⁷ and DrugCentral¹⁸ databases. On the other hand, T_{bio} (biology) refers to those proteins with known biological role and some evidences of linkage to a disease phenotype despite lacking an identified small molecule or approved drug with biological activity for them. Finally, T_{dark} (dark genome) assignments refer to the remaining proteins that have been manually-curated at the primary sequence level in UniProt,¹⁹ yet do not meet any of the criteria for T_{clin} , T_{chem} or T_{bio} . According to this, protein pairs targeted by dual agents can be classified into three categories, as follows:

- **$T_{\text{clin}}T_{\text{clin}}$ (clinic)**: target pair constituted by two T_{clin} proteins. They can be subclassified into four subcategories according to the MoA of the dual agent targeting them, if available:
 - **$T_{\text{clin}}T_{\text{clin}}|_{\text{MOA}}$ (clinic MOA)**, if both T_{clin} are targeted by the same DID_{MOA} and are annotated as the drug MoA;
 - **$T_{\text{clin}}T_{\text{clin}}|_{\text{MIX}}$ (clinic MIX)**, if both T_{clin} are targeted by the same DID_{MIX} and one of them is annotated as the drug MoA;
 - **$T_{\text{clin}}T_{\text{clin}}|_{\text{OFF}}$ (clinic OFF)**, if both T_{clin} are targeted by the same DID_{OFF} with none of them annotated as its MoA; and

- $T_{\text{clin}}T_{\text{clin|LIG}}$ (**clinic LIG**), if both T_{clin} are targeted by the same DIL.
- $T_{\text{clin}}T_{\text{chem}}$ (**mixed**): target pair constituted by one T_{clin} and one T_{chem} proteins. They can be subclassified into three subcategories according to the dual agent MoA when found:
 - $T_{\text{clin}}T_{\text{chem|MIX}}$ (**mixed MIX**), if both proteins are targeted by the same DID_{MIX} and the T_{clin} is annotated as the drug MoA;
 - $T_{\text{clin}}T_{\text{chem|OFF}}$ (**mixed OFF**), if both proteins are targeted by the same DID_{OFF} with none of them annotated as its MoA, and so considered as off-targets; and
 - $T_{\text{clin}}T_{\text{chem|LIG}}$ (**mixed LIG**), if both proteins are targeted by the same DIL.
- $T_{\text{chem}}T_{\text{chem}}$ (**chemistry**): target pair constituted by two T_{chem} proteins. They can be subclassified into two subcategories according to the dual agent:
 - $T_{\text{chem}}T_{\text{chem|OFF}}$ (**chemistry OFF**), if both T_{chem} are targeted by the same DID_{OFF} ; and
 - $T_{\text{chem}}T_{\text{chem|LIG}}$ (**chemistry LIG**), if both T_{chem} are targeted by the same DIL.

On the other hand, target pairs can also be classified according to the phylogenetic relation between their targets as follow:

- TT_{intra} (**intra-family**), if both targets are within the same protein family; and
- TT_{inter} (**inter-family**), if targets are classified in different protein families.

Map of dual agent vs target vs target triads

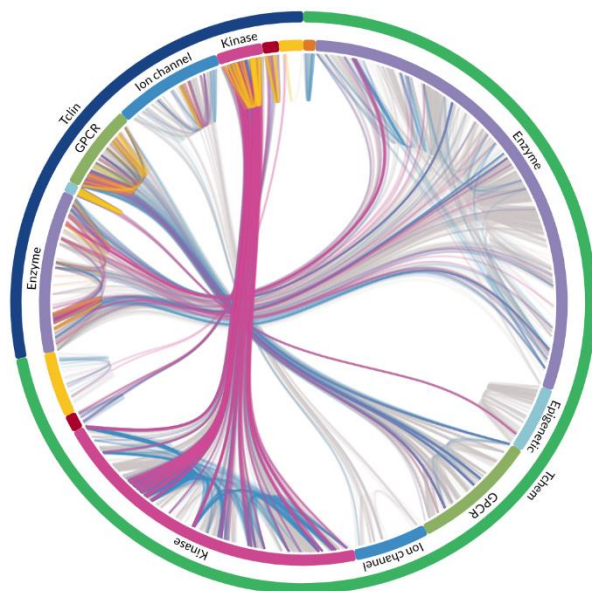
When applying the concepts defined to the T_{chem} and T_{clin} targets available at the Target Central Resource Database (TCRD)²⁰ (see Databases and Methods), a dataset of 134,270 dual agent *vs* target *vs* triads is generated.

Dual agents coverage. The triads identified are targeted by a total of 130 unique DID_{MOA} , 195 unique $DID_{\text{MIX}}/DID_{\text{OFF}}$, and 47,689 DILs. As expected, DID_{MOA} found are mainly anticancer therapeutics targeting multiple tyrosine-kinases; such as bosutinib, crizotinib, dasatinib and lapatinib; and antidepressants acting on multiple receptors of the same protein family; such as dexmedetomidine, cariprazine, butorphanol and bazedoxifene. They are represented in Figure 1 by the clusters of yellow connections accumulated at the T_{clin} section, which are specially enriched within kinases and G protein-coupled receptors (GPCRs) families. Our findings contradict the general viewpoint that approved dual-acting agents are the product of a rational drug design process.²¹ Instead, it is more likely that this was a result of serendipity as mentioned above.

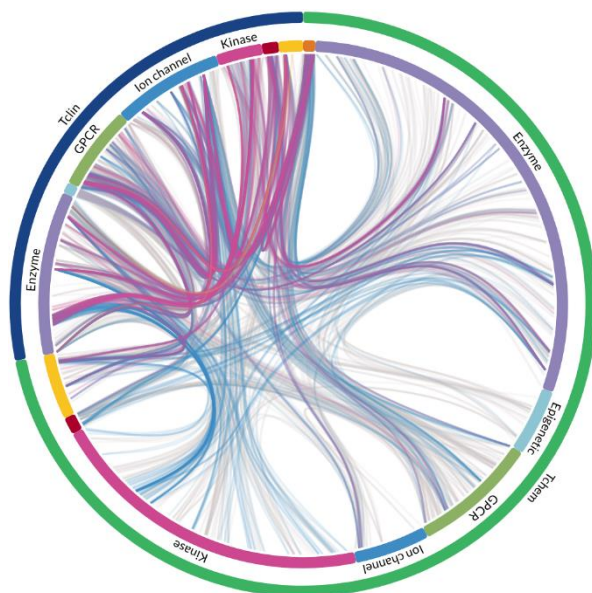
Target pairs coverage. The 134,270 triads identified can be collapsed into 9,549 unique target pairs. They are distributed in 2,717 (28%) $T_{\text{clin}}T_{\text{clin}}$, 4,009 (42%) $T_{\text{clin}}T_{\text{chem}}$ and 2,823 (30%) $T_{\text{chem}}T_{\text{chem}}$ (tab. 1, fig. 2). This means that almost three quarters of the target pairs found imply a target which is known to be clinically relevant. The $T_{\text{chem}}T_{\text{chem}}$ found correspond mainly to pairs of kinases and other enzymes targeted by tyrosine kinase inhibitors (TKI), so well known for their promiscuity.⁷ $T_{\text{clin}}T_{\text{clin}}$ pairs are mainly GPCRs (592), enzymes (599), kinases (429) and ion channels (193). This is not surprising, since these protein families have been intensively targeted by the pharmaceutical industry for decades due to their therapeutic interest in

cancer, channelopathies and neuropsychiatric disorders.¹⁶ In addition to the 6,842 (72%) intra-family pairs found, we identified 2,707 (28%) inter-family dyads, with GPCRs *vs* ion channels being the most populated subgroup. The majority of these TT_{inter} imply one or two safety-related targets and come from the testing of approved drugs in search for off-target undesirable adverse drug reactions.²² These safety-related targets include the potassium voltage-gated channel subfamily H member 2 (KCNH2 or hERG), whose blockade is associated with potentially fatal cardiac arrhythmias;²³ the muscarinic acetylcholine receptors, which have a fundamental role in physiology and should not unintendedly perturbed; the 5-hydroxytryptamine receptor 2B (5-HT_{2B}), associated with potential cardiac valvulopathy and pulmonary hypertension; and some tyrosine-protein kinases like Fyn, Lck, FLT3 and ABL1 that are regarded as ‘sentinel’ representatives of the adverse drug reactions-related kinases.²² The other non-safety-related inter-family target pairs imply in most of the cases a cytochrome P450, a carbonic anhydrase, a poly(ADP-ribose) polymerase (PARP), a cyclooxygenase, cholinesterase or acetylcholinesterase. In the *Supplementary Material* an exemplary list of intra-family and inter-family target pairs is provided.

Results



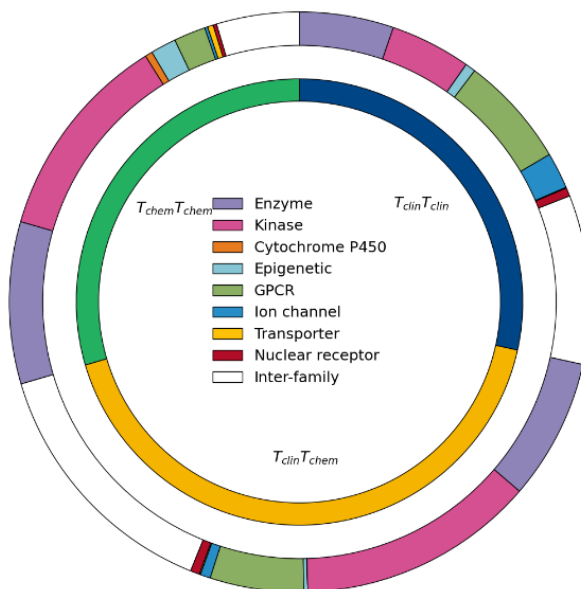
Intra-family target pairs



Inter-family target pairs

Figure 1. Map of the targetable human proteome targeted by dual agents. Nodes in the circles represent the initial T_{clin}/T_{chem} targets in the analysis. Those that are interconnected represent target pairs sharing a common dual agent.

a)



b)

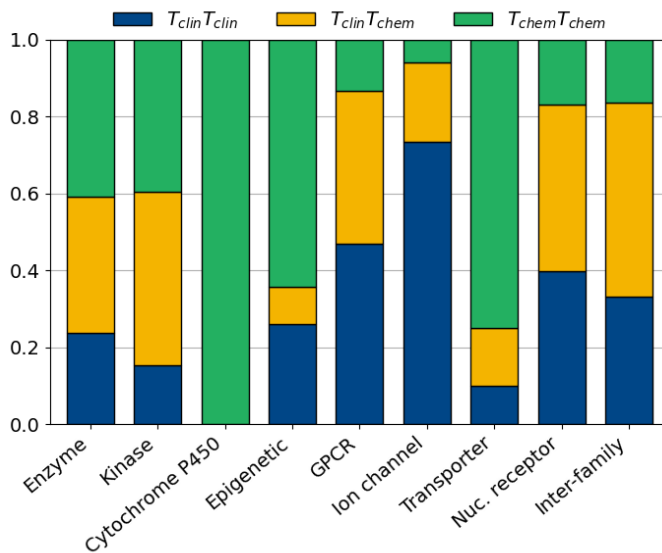


Figure 2. Distribution of 9,549 target pairs across protein families. **a)** Percentages of the target pairs categories. **b)** Target pairs categories across protein families. GPCR, G protein-coupled receptor.

Results

Table 1. Target pairs categories.

Protein family	Targets	T_{clin}T_{clin}	T_{clin}T_{chem}	T_{chem}T_{chem}
Enzymes	2,112	500 (24%)	750 (36%)	862 (41%)
Kinases	2,804	429 (15%)	1,264 (45%)	1,111 (40%)
Cytochrome P450s	39	0 (0%)	0 (0%)	39 (100%)
Epigenetic proteins	212	55 (26%)	21 (10%)	136 (64%)
GPCRs	1,259	592 (47%)	498 (40%)	169 (13%)
Ion channels	263	193 (73%)	54 (21%)	16 (6%)
Transporters	40	4 (10%)	6 (15%)	30 (75%)
Nuclear receptors	113	45 (40%)	49 (43%)	19 (17%)
Inter-family	2,707	899 (33%)	1,367 (50%)	441 (16%)
Total	9,549	2,717 (28%)	4,009 (42%)	2,823 (30%)

GPCR, G protein-coupled receptor.

Target pairs relative coverage. In the previous section, we analysed the amount of protein pairs found to be targeted by a common dual agent in terms of absolute coverage. Now, a complementary analysis in terms of relative coverage is carried out to put the absolute coverage figures in the context of the initial $T_{\text{clin}}+T_{\text{chem}}$ dataset. We calculated relative coverages across target pairs' categories and across target pairs' families by dividing absolute coverage values by the sum of all the theoretically possible target pairs that we could have at each group. For example, in an ideal case where a dual agent could be found for every target pair in the kinases family, based on the initial 425 (373+52) $T_{\text{clin}}+T_{\text{chem}}$ proteins (see Supplementary Material), the 2,804 target pairs found to be covered by a dual agent would represent a relative coverage equal to 3% ($2,804/(425*(425-1)/2)$). We are aware that 'not all target pair combinations will be accessible to a single agent with drug-like properties',¹ and that more realistic estimates of the amount of empirically possible targetable target pairs could be devised. Still this approach could shed some light on relative coverages at this point.

Results obtained from our target pairs' relative coverage analysis are shown in Table 2. According to it, clinical pairs (0.9%) are on average 8 times better covered than chemical ones (0.1%), which was expected given their known therapeutic utility. This contrast with absolute coverage results where clinical (29%, 2,717/9,549) and chemical (30%, 2,823/9,549) targets pairs showed almost equal values. This evidences the bias that analysing absolute coverages alone may introduce. Focusing on protein families, $T_{\text{clin}}T_{\text{clin}}$ epigenetic factors appear to be the most relatively covered family (83%). After $T_{\text{clin}}T_{\text{clin}}$ epigenetic factors, $T_{\text{chem}}T_{\text{chem}}$ cytochrome P450s have a relative coverage of 50%, stemming from compounds tested against different cytochrome P450s to assess metabolic liabilities. $T_{\text{clin}}T_{\text{clin}}$ kinases follow, with up to 32% relative coverage, which is explained by the large

Results

number of highly-promiscuous TKI having affinity for the ATP-binding site characteristic for protein-tyrosine kinases. Two other promiscuous families are nuclear receptors, with a $T_{\text{clin}}T_{\text{clin}}$ relative coverage equal to 29%; and GPCRs, with a $T_{\text{clin}}T_{\text{clin}}$ relative coverage of 13%. This is explained by the presence of highly conserved binding sites for monoamine neurotransmitters reception in most GPCRs and by the fact that they are the largest family of druggable targets in the human genome, with between 20% and 30% of approved drugs acting on them.¹⁶ On the contrary, families with the lowest relative coverages are enzymes, ion channels and transporters with values below 3%. Again, family-based relative coverage results contrast with absolute one since enzymes were the second most absolutely covered protein family (22%). Finally, as expected, relative coverages for inter-family pairs are below 1%, which further confirms that when two proteins are phylogenetically closer, it is easier to find a dual-acting agent.

Conclusions

The classical view of the compound *vs* target interactome as a 1-dimension map, might not be sufficient anymore to map the vast range of polypharmacological opportunities that might be enclosed in currently known chemical and biological spaces, and new terms and tools with a network-based perspective should be proposed. In this work, a novel ontology of terms to help map dual-pharmacological opportunities present in the human proteome has been proposed. When applied to compound *vs* target bioactivity data available in public sources, a set of almost 10 thousand protein pairs with a bioactive dual agent targeting them was

obtained. Almost three-fourths of these target pairs imply at least one mechanism-of-action target of an approved drug or/and imply two phylogenetically-related targets classified within the same protein family.

Proposed as a starting point for the analysis of dual-pharmacological opportunities, new layers of complexity could be added in the future to explore a wider range of polypharmacological opportunities.

	Absolute coverages		Relative coverages			
	T _T /9,549	T _{clin} T _{clin} /2,717	T _T	T _{clin} T _{clin}	T _{clin} T _{chem}	T _{chem} T _{chem}
Enzymes	22.1%	18.4%	0.8%	2.9%	0.7%	0.6%
Kinases	29.4%	15.8%	3.1%	32.4%	6.5%	1.6%
Cyt. P450s	0.4%	0.0%	50.0%	-	-	50.0%
Ep. proteins	2.2%	2.0%	5.5%	83.3%	2.3%	4.8%
GPCRs	13.2%	21.8%	24.9%	13.0%	3.7%	1.7%
Ion channels	2.8%	7.1%	1.2%	2.5%	0.5%	0.4%
Transporters	0.4%	0.1%	0.8%	1.1%	0.3%	1.1%
Nuclear rec.	1.2%	1.7%	17.0%	29.4%	14.3%	11.1%
Inter-family	28.3%	33.1%	0.2%	0.9%	0.3%	0.1%

Table 2. Target pairs absolute coverage *vs* relative coverage. Absolute coverages are calculated by dividing the total amount of target pairs and the total amount of T_{clin}T_{clin} pairs in each family by 9,549 and 2,717, respectively. Relative coverages are calculated by dividing the total amount of target pairs at each family and category by the sum of all theoretically possible targetable target pairs within each family and category.

Databases and Methods

Protein targets. The target development level (TDL) is a knowledge-based classification scheme, conceived as part of the Illuminating the Druggable genome (IDG) initiative, that groups proteins in four categories reflecting the depth of investigation from a clinical (T_{clin}), chemical (T_{chem}), therapeutical (T_{bio}), and biological (T_{dark}) standpoint. We downloaded the list of T_{clin} and T_{chem} proteins from the Target Central Resource Database v5.4.2exp and selected those targets classified by the Drug Target Ontology (DTO) as enzymes, kinases, cytochrome P450s, epigenetic proteins, G-protein coupled receptors (GPCRs), ion channels, transporters and nuclear receptors. See Supplementary Material for a summary of the initial dataset size.

Molecule vs target bioactivities. From ChEMBL v26, we downloaded bioactivities annotated to the T_{clin} and the T_{chem} in our list of targets, that fulfil the following conditions: (i) come from a binding ('B') or functional ('F') assay (ii) with a confidence score equal to 9, (iii) imply a small molecule, (iii) have an activity type like IC50, EC50, Ki or Kd and (iv) have a pChEMBL value equal or above the Oprea's *et al.* (2018)¹⁵ family-specific thresholds: ≤ 30 nM for kinases, ≤ 100 nM for GPCRs and nuclear receptors, ≤ 10 μM for ion channels and ≤ 1 μM for other target families. Using DrugCentral v2020 we annotated ChEMBL interactions linked by mechanism of action.

Dual agent vs target pair triads. We cross-matched the list of targets in order to obtain all the combinations of proteins sharing a common bioactive molecule (InChiKeys-based). This resulted in the final set of 134,270 dual

agent *vs* target *vs* triads. Triads coming from Staurosporine, the only pan-inhibitor found, were discarded to avoid bias in the analysis.

References

1. Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol.* **2008**, *4*(11), 682–690.
2. Anighoro, A.; Bajorath, J. & Rastelli, G. Polypharmacology: challenges and opportunities in drug discovery. *J Med Chem.* **2014**, *57*(19), 7874–7887.
3. Al-Lazikani, B.; Banerji, U. & Workman, P. Combinatorial drug therapy for cancer in the post-genomic era. *Nat Biotechnol.* **2012**, *30*(7), 679–692.
4. Gonzalez de Castro, D.; Clarke, P. A.; Al-Lazikani, B. & Workman P. Personalized cancer medicine: molecular diagnostics, predictive biomarkers, and drug resistance. *Clin Pharmacol Ther.* **2013**, *93*(3), 252–259.
5. Kitano, H. A robustness-based approach to systems-oriented drug design. *Nat Rev Drug Discov.* **2007**, *6*(3), 202–210.
6. Menden, M. P. *et al.* Community assessment to advance computational prediction of cancer drug combinations in a pharmacogenomic screen. *Nat Commun.* **2019**, *10*(1), 2674.
7. Knight, Z. A.; Lin, H. & Shokat, K. M. Targeting the cancer kinome through polypharmacology. *Nat Rev Cancer.* **2010**, *10*(2), 130–137.
8. Rivero-García, I. *et al.* Drug repurposing improves disease targeting 11-fold and can be augmented by network module targeting, applied to COVID-19. *Sci Rep.* **2021**, *11*(1), 20687.

Results

9. Patyar, S.; Prakash, A. & Medhi, B. Dual inhibition: a novel promising pharmacological approach for different disease conditions. *J Pharm Pharmacol.* **2011**, *63*(4), 459–471.
10. Keller, G.; Schafhausen, P. & Brummendorf, T. H. Bosutinib: a dual SRC/ABL kinase inhibitor for the treatment of chronic myeloid leukemia. *Expert Rev Hematol.* **2009**, *2*(5), 489–497.
11. Geyer, C. E. *et al.* Lapatinib plus capecitabine for HER2-positive advanced breast cancer. *N. Engl. J. Med.* **2006**, *355*, 2733–2743.
12. Jain R. Single-action versus dual-action antidepressants. *Prim Care Companion J Clin Psychiatry* **2004**, *6*(Suppl 1), 7–11.
13. Roth, B. L.; Sheffler, D. J. & Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat Rev Drug Discov.* **2004**, *3*(4), 353–359.
14. Paolini, G. V. *et al.* Global mapping of pharmacological space. *Nat Biotechnol.* **2006**, *24*(7), 805–815.
15. Fabian, M. A. *et al.* A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat Biotechnol.* **2005**, *23*(3), 329–336.
16. Oprea, T. *et al.* Unexplored therapeutic opportunities in the human genome. *Nat Rev Drug Discov.* **2018**, *17*, 317–332.
17. Mendez, D. *et al.* ChEMBL: Towards Direct Deposition of Bioassay Data. *Nucleic Acids Res.* **2019**, *47*, D930–D940.
18. Avram, S. *et al.* DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Res.* **2021**, *49*(D1), D1160–D1169.
19. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*(D1), D480–D489.
20. Sheils, T. *et al.*, TCRD and Pharos 2021: mining the human proteome for disease biology, *Nucl. Acids Res.* **2021**, *49*(D1), D1334–D1346.

21. Hopkins, A. L.; Mason, J. S. & Overington, J. P. Can we rationally design promiscuous drugs?. *Curr Opin Struct Biol.* **2006**, *16*(1), 127–136.
22. Bowes, J. *et al.* Reducing safety-related drug attrition: the use of in vitro pharmacological profiling. *Nat Rev Drug Discov.* **2012**, *11*(12), 909–922.
23. Redfern, W. S. *et al.* Relationships between preclinical cardiac electrophysiology, clinical QT interval prolongation and torsade de pointes for a broad range of drugs: evidence for a provisional safety margin in drug development. *Cardiovasc Res.* **2003**, *58*(1), 32–45.

Supplementary Material

Supplementary Table S1. Initial dataset of target proteins.

Protein family	Targets	T _{clin}	T _{chem}
Enzymes	740	187	553
Kinases	425	52	373
Cytochrome P450s	13	0	13
Epigenetic proteins	88	12	76
GPCRs	238	96	142
Ion channels	211	126	85
Transporters	101	27	74
Nuclear receptors	37	18	19
Total	1,853	518	1,335

Results

Supplementary Table S2. Examples of $T_{clin}T_{clin}$ pairs covered by a dual agent.

Intra-family TTs				
Target ₁ (gene)	Target ₂ (gene)	Dual agent	pAct ₁	pAct ₂
$T_{clin}T_{clin} MOA$				
Mast/stem cell growth factor receptor Kit (<i>KIT</i>)*	Platelet-derived growth factor receptor alpha (<i>PDGFR4</i>)*	Imatinib	7.84	7.85
Carbonic anhydrase 12 (<i>CA12</i>)*	Carbonic anhydrase 2 (<i>CA2</i>)*	Acetazolamide	8.23	7.86
$T_{clin}T_{clin} MIX$				
D(2) dopamine receptor (<i>DRD2</i>)*	Muscarinic acetylcholine receptor M1 (<i>CHRM1</i>)	Olanzapine	7.73	8.00
Estrogen receptor (<i>ESR1</i>)*	Progesterone receptor (<i>PGR</i>)	Fulvestrant	8.86	9.68
$T_{clin}T_{clin} OFF$				
Alpha-2C adrenergic receptor (<i>ADRA2C</i>)	5-hydroxytryptamine receptor 2C (<i>HTR2C</i>)	Amitriptyline	8.07	8.50
Fibroblast growth factor receptor 1 (<i>FGFR1</i>)	Vascular endothelial growth factor receptor 1 (<i>FLT1</i>)	Cabozantinib	7.95	7.92
$T_{clin}T_{clin} LIG$				
Sodium-dependent noradrenaline transporter (<i>SLC6A2</i>)	Sodium-dependent dopamine transporter (<i>SLC6A3</i>)	(R,S)-Indatraline	7.74	7.61
Mast/stem cell growth factor receptor Kit (<i>KIT</i>)	Platelet-derived growth factor receptor alpha (<i>PDGFR4</i>)	Cediranib	9.45	9.39
D(2) dopamine receptor (<i>DRD2</i>)	5-hydroxytryptamine receptor 2A (<i>HTR2A</i>)	Blonanserin	9.85	9.09
Histone deacetylase 4 (<i>HDAC4</i>)	Histone deacetylase 11 (<i>HDAC11</i>)	CHEMBL42278 98	8.12	8.12

Supplementary Table S2. (continued). *Target annotated as the MoA of the dual agent.

Inter-family TTs				
Target ₁ (gene)	Target ₂ (gene)	Dual agent	pAct ₁	pAct ₂
T_{clin}T_{clin} MOA				
Sodium-dependent noradrenaline transporter (<i>SLC6A2</i>)*	Histamine H1 receptor (<i>HRH1</i>)*	Doxepin	7.54	9.84
5-hydroxytryptamine receptor 2C (<i>HTR2C</i>)*	Sodium-dependent serotonin transporter (<i>SLC6A4</i>)*	Trazodone	7.33	6.99
T_{clin}T_{clin} MIX				
Sodium-dependent serotonin transporter (<i>SLC6A4</i>)*	Muscarinic acetylcholine receptor M4 (<i>CHRM4</i>)	Amitriptyline	8.07	9.13
5-hydroxytryptamine receptor 2A (<i>HTR2A</i>)*	Sodium-dependent serotonin transporter (<i>SLC6A4</i>)	Ziprasidone	9.44	7.11
Mu-type opioid receptor (<i>OPRM1</i>)*	Sodium channel protein type 5 subunit alpha (<i>SCN5A</i>)	Loperamide	9.28	6.62
T_{clin}T_{clin} OFF				
5-hydroxytryptamine receptor 2A (<i>HTR2A</i>)	Sodium-dependent serotonin transporter (<i>SLC6A4</i>)	Chlorpromazine	8.46	7.44
Rod cyclic nucleotide phosphodiesterase subunit alpha (<i>PDE6A</i>)	Adenosine receptor A2a (<i>ADORA2A</i>)	Sildenafil	8.1	6.78
Acetylcholinesterase (<i>ACHE</i>)	5-hydroxytryptamine receptor 2C (<i>HTR2C</i>)	Fluoxetine	6.89	6.92
T_{clin}T_{clin} LIG				
Delta-type opioid receptor (<i>OPRD1</i>)	Gamma-aminobutyric acid receptor subunit alpha-2 (<i>GABRA2</i>)	Amentoflavone	7.44	8.22

Part IV: Discussion

Patents as a source of novel chemical space

The new generation of patent chemical repositories obtained by means of artificial intelligence (AI) algorithms⁸⁸ is disclosing a whole novel chemical space of biological relevance. The main limitation is their incapability to discriminate between compounds of pharmacological relevance and other starting materials and intermediate products also appearing in the patent document.¹⁸ To address this situation, in Chapter III.1, I introduce a new unsupervised filtering protocol able to automatically select the core chemical structures best representing a congeneric series of pharmacologically relevant molecules in a patent. Validated with the manually curated ChEMBL¹⁴ patents, and applied then to the recently released SureChEMBL patent chemical database,¹⁸ a final set of ~6 million molecules conforming congeneric chemical series in ~0.2 million patents is disclosed. Open access to this filtered SureChEMBL version enriched with molecules of pharmacological relevance, named SureChEMBLccs, is available for download at the EMBL-EBI web site.⁸⁹ The protocol presented could be used to generate regular updates of SureChEMBLccs and be extended to other AI-generated chemical patent databases, such as Google Patents,⁹⁰ for a greater coverage of patent-derived pharmacologically-relevant chemical space.

Complementary, I believe that future research in this line should be devoted to devising approaches to illuminate the biological space contained in patent documents as well. Protein targets and disease indications for the claimed compounds can also be found between patent documents' lines, so mapping patent-derived chemical information to patent-derived biological information would help us close the so sought-after circle of drug *vs* target *vs* disease.

The chemical scope of patent applications

The sets of molecules claimed in patents can be chemically defined by a Markush structure, which is an abstract representation of the region of the chemical space protected by the patent. In Chapter III.2, I have demonstrated that this patent claimed chemical region tend to be, on the one hand, narrow enough to retain only chemical series of compounds acting on the same mechanism-of-action target(s) or having similar cellular phenotype(s), while on the other hand, wide enough to cover as much as possible compounds with these characteristics to ensure the exclusivity of the invention.

This same trend is observed both in patents deposited in manually-curated databases like ChEMBL and in automatically-generated ones like SureChEMBL. However the second ones show greater compounds coverage compared to the first ones, without losing their high molecular congenericity degree. What is left to know is whether this coverage increase occurs only in terms of quantity or, what is more interesting, in terms of diversity. If it is the second, facilitating early patent access and developing reliable and efficient tools to explore their content will be extremely valuable for advancing in the discovery of medicinal chemistry solutions.

The similarity-based approach introduced, was applied in this chapter to assess the molecular congenericity of collections of compounds claimed by patents. However, it could indistinctly serve to describe the molecular congenericity degree of any collection of molecules. Moreover, its capability to summarize compounds sets similarities in just a few descriptors allows for the easy comparison across different collections of molecules.

Illuminating the chemical space of untargeted proteins

The IDG initiative estimates that 10% of the proteins are mechanism-of-action targets of an approved drug or have at least one bioactive ligand deposited in public sources. The remaining 90% of the human proteome is composed by chemically neglected proteins that nonetheless have well established implications in biological processes or their primary sequences is all what is currently known.⁷⁰ Some of the reasons hindering the identification of bioactive ligands for understudied proteins include the difficulty of identifying ligand binding sites in ligand-independent orphan targets,^{91,92} the absence of optimal assays to detect the activation of receptors with atypical coupling,^{93,94} and the absence of protein family members with already known bioactive ligands.⁶⁸ All these aspects, added to increased high-throughput screening costs and more stringent safety regulations, justify that research in this field prefers investing on generating novel chemical series for already targeted proteins over initiating high risk projects on yet untargeted ones.^{68,95}

Following the paradigm that similar proteins are likely to interact with similar ligands, in Chapter III.3 I applied a similar concept to that of the core structures described in Chapter III.1, in order to identify those core scaffolds that are more enriched within families of phylogenetically-related proteins and that, when expanded, are highly probable of containing active small molecules for untargeted proteins included in the family. When applied to a set of 128 untargeted proteins, these privileged core scaffolds allowed for extracting a priority list of commercially available small molecules with a minimum success rate of delivering active ligands for the untargeted proteins equal to 32%.

A novel ontology to help DADs discovery

As the knowledge on the molecular mechanisms behind drugs action increases, it is becoming more evident that many drugs approved in the past as single-action ones, in fact owe their efficacy to their multi-action behaviour.⁹⁶ This is prompting a shift in the perception of drug off-targets effects in the medicinal chemistry community. From being considered as prone to safety issues, undesirable and necessary to be removed, they are progressively been seen as promising opportunities for multi-target drugs' discovery.⁹⁶

The classical view of the compound *vs* target interactome as a 1-dimension map, might not be sufficient anymore to map the vast range of polypharmacological opportunities that might be enclosed in currently known chemical and biological spaces, and new terms and tools with a network-based perspective should be proposed. To contribute, in Chapter III.4, I present a novel ontology of terms to help describe and catalogue the current human proteome that might be targetable by a dual-acting agent. A list of target pairs with a dual-acting agents of therapeutic interest is also provided for future research.

The analysis is initially proposed for the lowest level of polypharmacology (dual-pharmacology) since dual-action drugs with efficacy in the treatment of CNS disorders and different types of cancer have already been approved.⁸¹⁻⁸⁷ However, new layers of complexity could be added in the future to explore a wider range of polypharmacological opportunities.

Part V: Conclusions

The main contributions of this Thesis can be summarized as follows:

- i) A new unsupervised protocol has been designed to automatically identify the core chemical structures best representing the congeneric series of pharmacologically-relevant molecules in patents.
- ii) The protocol has been applied to generate a new filtered version of SureChEMBL database enriched with molecules of pharmacological relevance, which is available for download at an EMBL ftp site under the name of SureChEMBLccs.
- iii) This new SureChEMBLccs database gives direct access to a pharmacologically-relevant chemical space of ~6 million molecules selected from ~0.2 million US pharmacological patents. This could be enriched with future updates of the database and by extending the application of the protocol to other chemical patent-derived databases.
- iv) A similarity-based method to analyse quantitatively the degree of structural congenericity of collections of compounds has been introduced.
- v) The method has been applied to evaluate the degree of congenericity of claimed compounds in patent applications, demonstrating its efficacy to differentiate between patents exemplifying highly congeneric compounds of a structurally compact and well defined chemical series, from patents containing a more diverse set of compounds around a more vaguely described patent claim.

Conclusions

- vi) The method has been applied to evaluate the congenericity of patent molecular compositions coming from different chemical sources; namely SureChEMBL, SureChEMBLccs and ChEMBL; concluding that filtered molecules in SureChEMBLccs show overlapping congenericities with the manually curated sets in ChEMBL.
- vii) A new computational protocol to identify those core scaffolds that are more enriched within families of phylogenetically-related proteins has been proposed to help illuminate the biologically-active chemical space of yet untargeted proteins included in the families.
- viii) The protocol has been applied to a set of 128 proteins that still lack a bioactive ligand in public sources. The core scaffolds obtained have allowed for extracting a priority list of commercially available small molecules with a minimum success rate of delivering active ligands for the untargeted proteins equal to 32%.
- ix) A novel ontology to help map dual-pharmacological opportunities present in the human proteome has been proposed, revealing that a dual-acting agent with biologically-relevant activity exists for almost 10 thousand protein target pairs.
- x) Almost three-fourths of these target pairs for which a dual-acting agent was identified involve phylogenetically-related targets classified within the same protein family and have at least one mechanism-of-action target of an approved drug. The remaining one-fourth is composed of pairs of proteins coming from different families yet neglected as mechanism-of-action drug targets despite having known bioactive ligands deposited in public sources

Part VI: References

1. Jackson, B. How many stars are there in space?. *The Conversation* **2021**, September. Retrieved from <https://theconversation.com/how-many-stars-are-there-in-space-165370> (last accessed on December 13th, 2021).
2. Reymon, J. L. & Awale, M. Exploring chemical space for drug discovery using the chemical universe database. *ACS Chem Neurosci.* **2012**, *3*(9), 649–57.
3. Bohacek, R. S.; McMartin, C. & Guida, W. C. The art and practice of structure-based drug design: a molecular modelling perspective. *Med. Res. Rev.* **1996**, *16*, 3–50.
4. NASA; ESA & Hubble Heritage Team (STScI/AURA). The pillars of creation. *NASA* **2018**, February. Retrieved from <https://www.nasa.gov/image-feature/the-pillars-of-creation> (last accessed on December 13th, 2021).
5. Drews, J. Drug discovery: A historical perspective. *Science* **2000**, *287*(5460), 1960–1964.
6. Macarron, R. *et al.* Impact of high-throughput screening in biomedical research. *Nat Rev Drug Discov* **2011**, *10*(3), 188–95.
7. Dobson, C. M. Chemical space and biology. *Nature* **2004**, *432*(7019), 824–8.
8. Dolle, R. E. Comprehensive survey of combinatorial library synthesis: 2002. *J. Comb. Chem.* **2003**, *5*, 693–753.
9. Hastings, J. *et al.* ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.* **2016**, *44*(D1), D1214–9.
10. Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **2021**, *49*(D1), D1388–D1395.
11. Editorial: ChemSpider--a tool for natural products research. *Nat Prod Rep.* **2015**, *32*(8), 1163–4.

References

12. Bobrowski, T. M.; Korn, D. R.; Muratov, E. N. & Tropsha, A. ZINC express: a virtual assistant for purchasing compounds annotated in the ZINC database. *J Chem Inf Model.* **2021**, *61*(3), 1033–1036.
13. Gilson, M. K. *et al.* BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **2016**, *44*(D1), D1045–53.
14. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **2019**, *47*(D1), D930–D940.
15. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **2018**, *46*(D1), D1074–D1082.
16. Avram, S. *et al.* DrugCentral 2021 supports drug discovery and repositioning. *Nucleic Acids Res.* **2021**, *49*(D1), D1160–D1169.
17. Harding, S. D. *et al.* The IUPHAR/BPS guide to PHARMACOLOGY in 2022: curating pharmacology for COVID-19, malaria and antibacterials, *Nucleic Acids Res.* **2021**, gkab1010.
18. Papadatos, G. *et al.* SureChEMBL: a large-scale, chemically annotated patent document database. *Nucleic Acids Res.* **2016**, *44*(D1), D1220–8.
19. Lipinski, C. A. & Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **2004**, *432*(7019), 855–61.
20. Oprea, T. I. & Gottfries, J. Chemography: the art of navigating in chemical space. *J Comb Chem.* **2001**, *3*(2), 157–66.
21. Pearlman, R. S. & Smith, K. Novel software tools for chemical diversity. *Perspectives in Drug Discovery and Design* **1998**, *9*, 339–353.
22. Lipinski, C. A.; Lombardo, F.; Dominy, B. W. & Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv Drug Deliv Rev.* **2001**, *46*(1–3), 3–26.

23. Leeson, P. D. & Springthorpe, B. The influence of drug-like concepts on decision-making in medicinal chemistry. *Nature reviews. Drug discovery* **2007**, *6*(11), 881–890.
24. Evans, B. E. *et al.* Methods for drug discovery: development of potent, selective, orally effective cholecystokinin antagonists. *Journal of medicinal chemistry* **1988**, *31*(12), 2235–2246.
25. Patchett, A. A. & Nargund, R. P. Chapter 26. Privileged structures — An update **1999**, *35*, 289–298.
26. Bemis, G. W. & Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *Journal of medicinal chemistry* **1996**, *39*(15), 2887–2893.
27. Kruger, F.; Fechner, N. & Stiefl, N. Automated Identification of Chemical Series: Classifying like a Medicinal Chemist. *Journal of chemical information and modelling* **2020**, *60*(6), 2888–2902.
28. de la Vega de León, A.; Hu, Y. & Bajorath, J. Systematic identification of matching molecular series and mapping of screening hits. *Molecular informatics* **2014**, *33*(4), 257–263.
29. Klabunde, T. & Hessler, G. Drug design strategies for targeting G-protein-coupled receptors. *Chembiochem.* **2002**, *3*(10), 928–944.
30. Gregori-Puigjané, E. & Mestres, J. Coverage and bias in chemical library design. *Curr Opin Chem Biol.* **2008**, *12*(3), 359–365.
31. Bregonje, M. Patents: A unique source for scientific technical information in chemistry related industry?. *World Patent Information* **2005**, *27*(4), 309–315.
32. Southan, C. Expanding opportunities for mining bioactive chemistry from patents. *Drug Discov Today Technol.* **2015**, *14*, 3–9.
33. Digital Science transfers SureChem patent chemistry data to EMBL-EBI. *EMBL-EBI Press Release* **2013**, December. Retrieved from

References

- <http://www.ebi.ac.uk/about/news/press-releases/SureChEMBL> (last accessed on December 13th, 2021).
34. Senger, S. Assessment of the significance of patent-derived information for the early identification of compound-target interaction hypotheses. *J Cheminform.* **2017**, 9(1), 26.
 35. Heifets, A. & Jurisica, I. SCRIpDB: a portal for easy access to syntheses, chemicals and reactions in patents. *Nucleic Acids Res.* **2012**, 40(D), D428–D433.
 36. Apodaca, R. L. IBM donates large collection of patent chemical structures to NIH/PubChem. *Depth-First* **2011**, December. Retrieved from <https://depth-first.com/articles/2011/12/15/ibm-donates-large-collection-of-patent-chemical-structures-to-nih-pubchem/> (last accessed on December 13th, 2021).
 37. Chambers, J. *et al.* UniChem: extension of InChI-based compound mapping to salt, connectivity and stereochemistry layers. *J Cheminform.* **2014**, 6(1), 43.
 38. Heller, S. R.; McNaught, A.; Pletnev, I.; Stein, S. & Tchekhovskoi, D. InChI, the IUPAC International Chemical Identifier. *Journal of Cheminformatics* **2015**, 7, 23.
 39. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, 28(1), 31–36.
 40. Weininger, D.; Weininger, A. & Weininger, J. L. SMILES. 2. Algorithm for generation of unique SMILES notation. *J. Chem. Inf. Comput. Sci.* **1989**, 29(2), 97–101.
 41. Weininger, D. SMILES. 3. DEPICT. Graphical depiction of chemical structures. *J. Chem. Inf. Comput. Sci.* **1990**, 30(3), 237–243.

42. 3. SMILES - A simplified chemical language. *DAYLIGHT Chemical Information Systems, Inc.* **2019**, May. Retrieved from <https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html> (last accessed on December 13th, 2021).
43. 4. SMARTS- A language for describing molecular patterns. *DAYLIGHT Chemical Information Systems, Inc.* **2019**, May. Retrieved from <https://www.daylight.com/dayhtml/doc/theory/theory.smiles.html> (last accessed on December 13th, 2021).
44. Dalby, A. *et al.* Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. *J. Chem. Inf. Model.* **1992**, 32(3), 244.
45. van de Waterbeemd, H. *et al.* Glossary of terms used in computational drug design (IUPAC Recommendations 1997). *Pure and Applied Chemistry* **1997**, 69(5), 1137–1152.
46. Molecular fingerprints and similarity searching. *Open Babel v2.3.1* **2011**. Retrieved from <https://openbabel.org/docs/dev/Fingerprints/intro.html> (last accessed on December 13th, 2021).
47. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, 3, 33.
48. Carhart, R. E.; Smith, D. H. & Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: Definition and applications. *J. Chem. Inf. Comp. Sci.* **1985**, 25, 64–73.
49. Nilakantan, R.; Bauman N.; Dixon, J. S. & Venkataraghavan, R. Topological torsion: A new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comp. Sci.* **1987**, 27, 82–5.

References

50. Rogers, D. & Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model.* **2010**, *50*, 742–54.
51. Apodaca R. L. Computing extended connectivity fingerprints. *Depth-First* **2019**, January. Retrieved from <https://depth-first.com/articles/2019/01/11/extended-connectivity-fingerprints/> (last accessed on December 13th, 2021).
52. Maggiora, G.; Vogt, M.; Stumpfe, D. & Bajorath, J. Molecular similarity in medicinal chemistry. *J. Med. Chem.* **2014**, *57*(8), 3186–3204.
53. Mestres, J. & Maggiora, G. M. Putting molecular similarity into context: asymmetric indices for field-based similarity measures. *Journal of mathematical chemistry* **2006**, *39*(1), 107–118.
54. O'Boyle, N. M. *et al.* Open Babel: An open chemical toolbox. *J. Cheminf.* **2011**, *3*, 33.
55. ChemAxon: Software solutions and services for chemistry and biology. <https://chemaxon.com/> (last accessed on December 13th, 2021).
56. Landrum, G. A. RDKit: Open-source cheminformatics software, version 2017.09.1; <http://www.rdkit.org> (last accessed on December 13th, 2021).
57. Venter, J. C. *et al.* The sequence of the human genome. *Science* **2001**, *291*(5507), 1304–1351.
58. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **2001**, *409*, 860–921.
59. Zahn, L. M. The human genome: A research ultramarathon. *Science* **2021**, *373*(6562), 1458–1459.
60. Gates, A. J.; Gysi, D. M.; Kellis, M. & Barabási, A. L. A wealth of discovery built on the Human Genome Project - by the numbers. *Nature* **2021**, *590*(7845), 212–215.

61. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*(D1), D480–D489.
62. Burley S. K. *et al.* RCSB Protein Data Bank: Celebrating 50 years of the PDB with new tools for understanding and visualizing biological macromolecules in 3D. *Protein Sci.* **2021**, 10.1002/pro.4213. doi:10.1002/pro.4213.
63. DeepMind and EMBL release the most complete database of predicted 3D structures of human proteins. *EMBL-EBI* **2021**, July. Retrieved from <https://www.ebi.ac.uk/about/news/press-releases/alphafold-database-launch> (last accessed on December 13th, 2021).
64. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589.
65. Protein classification. *EMBL-EBI*. Retrieved from <https://www.ebi.ac.uk/training/online/courses/protein-classification-intro-ebi-resources/protein-classification/what-are-protein-families/> (last accessed on December 13th, 2021).
66. Barrett, A. J. Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB). Enzyme nomenclature. Recommendations 1992. Supplement 4: corrections and additions (1997). *Eur J Biochem.* **1997**, *250*(1), 1–6.
67. Enzyme Commission Number. *Abnova*. Retrieved from <https://www.abnova.com/enzyme-commission-number> (last accessed on December 13th, 2021).
68. Laschet, C.; Dupuis, N. & Hanson, J. The G protein-coupled receptors deorphanization landscape. *Biochem Pharmacol.* **2018**, *153*, 62–74.
69. Klabunde, T. Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br J Pharmacol.* **2007**, *152*(1), 5–7.

References

70. Oprea, T. *et al.* Unexplored therapeutic opportunities in the human genome. *Nat Rev Drug Discov.* **2018**, *17*, 317–332.
71. Sheils, T. *et al.*, TCRD and Pharos 2021: mining the human proteome for disease biology, *Nucl. Acids Res.* **2021**, *49*(D1), D1334–D1346.
72. Hopkins, A. L. Network pharmacology: the next paradigm in drug discovery. *Nat Chem Biol.* **2008**, *4*(11), 682–690.
73. Sams-Dodd, F. Target-based drug discovery: is something wrong?. *Drug Discov Today* **2005**, *10*(2), 139–147.
74. Kaufmann, S. H. Paul Ehrlich: founder of chemotherapy. *Nat Rev Drug Discov.* **2008**, *7*(5), 373.
75. Kola, I. & Landis, J. Can the pharmaceutical industry reduce attrition rates?. *Nat Rev Drug Discov.* **2004**, *3*(8), 711–715.
76. Austin, C. P. *et al.* The knockout mouse project. *Nat Genet.* **2004**, *36*(9), 921–924.
77. Winzeler, E. A. *et al.* Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **1999**, *285*(5429), 901–906.
78. Giaever, G. *et al.* Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **2002**, *418*(6896), 387–391.
79. Kitano, H. A robustness-based approach to systems-oriented drug design. *Nat Rev Drug Discov.* **2007**, *6*(3), 202–210.
80. Knight, Z. A.; Lin, H. & Shokat, K. M. Targeting the cancer kinome through polypharmacology. *Nat Rev Cancer* **2010**, *10*(2), 130–137.
81. Patyar, S.; Prakash, A. & Medhi, B. Dual inhibition: a novel promising pharmacological approach for different disease conditions. *J Pharm Pharmacol.* **2011**, *63*(4), 459–471.

82. Stanković, T. *et al.* Dual Inhibitors as a New Challenge for Cancer Multidrug Resistance Treatment. *Curr Med Chem.* **2019**, *26*(33), 6074–6106.
83. Rahn, K. H. Clinical experience with dual-acting drugs in hypertension. *Clin Investig.* **1992**, *70*(Suppl 1), S39–S42.
84. Jain R. Single-action versus dual-action antidepressants. *Prim Care Companion J Clin Psychiatry* **2004**, *6*(Suppl 1), 7–11.
85. Keller, G.; Schafhausen, P. & Brummendorf, T. H. Bosutinib: a dual SRC/ABL kinase inhibitor for the treatment of chronic myeloid leukemia. *Expert Rev Hematol.* **2009**, *2*(5), 489–497.
86. Geyer, C. E. *et al.* Lapatinib plus capecitabine for HER2-positive advanced breast cancer. *N. Engl. J. Med.* **2006**, *355*, 2733–2743.
87. Roth, B. L.; Sheffler, D. J. & Kroeze, W. K. Magic shotguns versus magic bullets: selectively non-selective drugs for mood disorders and schizophrenia. *Nat Rev Drug Discov.* **2004**, *3*(4), 353–359.
88. Staker, J.; Marshall, K.; Abel, R. & McQuaw, C. M. Molecular structure extraction from documents using deep learning. *J. Chem. Inf. Model.* **2019**, *59*, 1017–1029.
89. SureChEMBLccs, **2021**.
<ftp://ftp.ebi.ac.uk/pub/databases/chembl/SureChEMBLccs> (last accessed on January 28th, 2021).
90. Google Patents, **2006**. <https://patents.google.com> (last accessed on December 13th, 2021).
91. Davenport, A. P. *et al.* International Union of Basic and Clinical Pharmacology. LXXXVIII. G protein-coupled receptor list: recommendations for new pairings with cognate ligands. *Pharmacol Rev.* **2013**, *65*(3), 967–986.

References

92. Levoye, A. *et al.* Do orphan G-protein-coupled receptors have ligand-independent functions? New insights from receptor heterodimers. *EMBO Rep.* **2006**, 7(11), 1094–1098.
93. Rajagopal, S. *et al.* Beta-arrestin- but not G protein-mediated signaling by the "decoy" receptor CXCR7. *Proc Natl Acad Sci U S A.* **2010**, 107(2), 628–632.
94. Okinaga, S *et al.* C5L2, a nonsignaling C5A binding protein. *Biochemistry.* **2003**, 42(31), 9406–9415.
95. Tunaru, S. Strategies for G-protein coupled receptor deorphanization. *Molecular Life.* **2017**, 1(1), 71–79.
96. Rivero-García, I. *et al.* Drug repurposing improves disease targeting 11-fold and can be augmented by network module targeting, applied to COVID-19. *Sci Rep.* **2021**, 11(1), 20687.

Appendix

Poster presented at the Symposium to Celebrate 10 Years of the ChEMBL Database, 2019, Hinxton (UK). It summarizes the article in Chapter III.1.

