



Funded by the Seventh Framework  
Programme of the European Union



Project full title:  
**Hepatic and Cardiac Toxicity Systems modelling**

Project acronym:  
**HeCaTos**

Collaborative project  
HEALTH.2013.1.3.-1:  
Modelling toxic response in case studies for predictive human safety assessment

**FP7-HEALTH-2013-INNOVATION-1-602156-HeCaTos**

**Deliverable Report D1.8:**  
**Report on the approach of bioassay pooling to improve model training**

*Edited by*  
*Nicolas Bosc and Fiona M I Hunter, EMBL-EBI*  
Work package 1

Due date of deliverable: M48  
Actual submission date: 30<sup>th</sup> September 2017

Start date of project: October, 2013

Duration: 60 months

**Maastricht University (UM)**

Project co-funded by the European Commission within the 7th Framework Programme (2013-2018)		
Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission	
CO	Confidential, only for members of the consortium (including the Commission Services)	

## Contributions to deliverable - Internal review procedure

Deliverable produced by:	Date:
Fiona MI Hunter - Partner EMBL	September 2017
Nicolas Bosc - Partner EMBL	September 2017
Deliverable internally reviewed by:	Date:
Anne Hersey - Partner EMBL	September 2017
Jos Kleinjans - Partner UM	September 2017

## Contents

<b>Publishable Summary .....</b>	<b>3</b>
<b>Objectives .....</b>	<b>3</b>
<b>1. Introduction .....</b>	<b>3</b>
<b>2. Selection of data from the ChEMBL database .....</b>	<b>4</b>
2.1. Selection of assays .....	5
2.2.1 Binding assay selection .....	5
2.1.2 <i>In vitro</i> functional assays.....	5
2.1.3 <i>In vivo</i> assays.....	5
2.1.4 Summary of selected assays .....	6
2.2. Selection of active compounds .....	6
2.3. Data selection expansion through molecular similarity.....	8
<b>3. Annotation of selected data and its representation as a graph database .....</b>	<b>10</b>
3.1. Annotation of the protein targets .....	10
3.2. Annotation of <i>in vivo</i> assays .....	13
3.3. Representation of the selected, annotated data as a graph database.....	15
3.4. Linking a protein target to its <i>in vivo</i> phenotypic outcome.....	16
<b>4. Results .....</b>	<b>19</b>
4.1. Using the graph database to investigate HeCaToS drugs.....	19
4.2. Similarity based cascades .....	23
4.3. Known toxicity mechanisms .....	23
4.4. Linking a protein target to its <i>in vivo</i> phenotypic outcome using a Spearman rank correlation between common compounds.....	24
<b>5. Difficulties .....</b>	<b>24</b>
<b>References .....</b>	<b>27</b>
<b>ANNEX 1: Selection of protein targets.....</b>	<b>28</b>
<b>Annex.....</b>	<b>34</b>
<b>References .....</b>	<b>46</b>

## Publishable Summary

The report describes the annotation of assays from the ChEMBL database to enable similar categories of data to be grouped together (bioassay pooling). Further, the annotated assays (and other associated data) were represented as a graph database because this provides improved flexibility and ease of use to investigate related data types. The graph database was used to investigate related assays and assess whether this approach can show promise as an *in silico* tool to enhance our knowledge of heart and liver toxicity. It was found that connected therapeutic pathways can be created from binding interactions at the protein level to phenotypic outcomes in whole animals.

## Objectives

The work presented in this report fulfils HeCaToS task T1.10: 'Develop approaches to extend scope of models through data pooling (cell lines, species, target families, etc). This report describes the work carried out to:

- Annotate assays described by the ChEMBL database in order to improve the database organisation so that similar categories of data can be grouped together (i.e. data pooling);
- Represent the annotated assays (and other associated data) as a graph database because this provides improved flexibility and ease of use to investigate related data types;
- Apply the graph database to investigate related assays and assess whether this approach can show promise as an *in silico* tool to enhance knowledge of heart and liver toxicity and their relationship to protein targets, based on existing data that is contained within the ChEMBL database.

## 1. Introduction

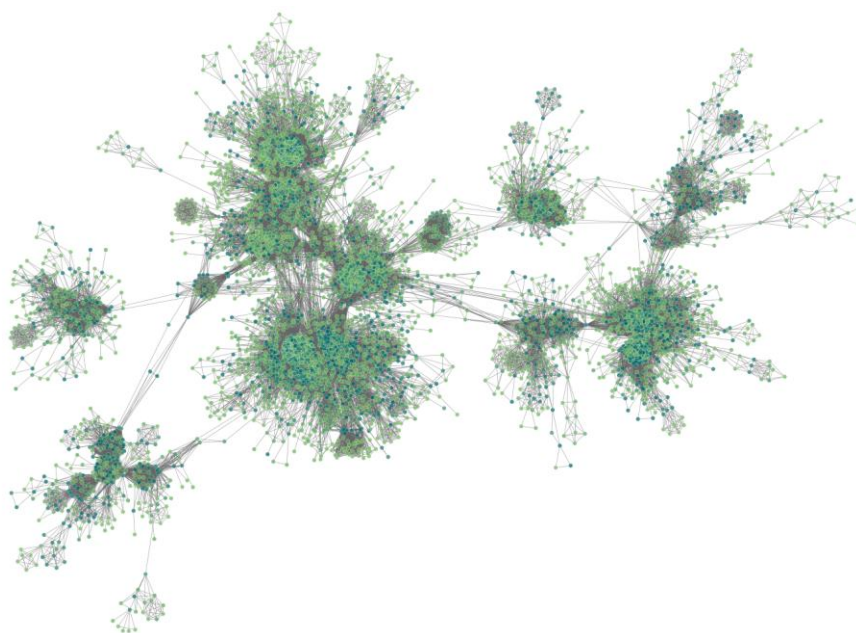
ChEMBL<sup>1</sup> (<https://www.ebi.ac.uk/chembl>) is a freely accessible database that provides information on more than a million high-quality, curated biological assays and on the bioactivity values measured between a wide variety of targets and compounds, as well as other information for e.g. approved drugs. For this work, the binding and functional assay data were considered. They represent more than 60 % of the ChEMBL database (version 22) and have been grouped as follows:

- Binding assays are considered to have been performed on a single protein in order to assess their interaction with a small molecule;
- Functional assays have been split in two categories depending on their target:
  - o An *in vivo* assay has been performed on a whole, living animal;
  - o An *in vitro* assay has been performed on isolated cells, cell-lines, isolated organs or isolated tissues. This group consists of all functional assays that have not been performed on whole, living animals (i.e. not *in vivo*).

Therefore, a binding assay considers a relatively simple biological system, in comparison to an *in vitro* functional assay for e.g. an isolated tissue, or the high complexity of an *in vivo* biological system for a whole animal. The ChEMBL database contains information on compounds that may have been tested in many different types of assay, and therefore it is possible to study their effect on biological systems at different levels of complexity. Hence, the assays described in the ChEMBL database can be linked if an individual compound can be shown to (i) actively inhibit an individual protein target and (ii) actively inhibit a cell-line or tissue, or organ, and (iii) show a phenotypic outcome in a whole animal such as a rat.

The ChEMBL database is a traditional relational database that stores data in a highly structured way, using predetermined entities across multiple tables. While this data management structure is excellent for maintaining large volumes of associated data types, it has limited flexibility to interrogate the data in a novel way without extensive effort. An alternative approach is to represent the same data as a graph database (Figure 1), which requires connections (edges) to be established

between entities (nodes). The result is an approach that is much simpler, and more expressive than the traditional relational database because it allows the data to be queried in multiple different ways to answer different scientific questions. For example, in this report the graph database has been used to connect similar assays that share the same active compound which means that relationships between compounds that bind to protein targets and also affect whole cells or organisms can be identified.



**Figure 1:** Example graph database showing nodes and edges to represent data

An alternative approach that will be considered for the next phase of work is to use the annotated assays (and their associated data) as a training dataset for future modelling needs. Indeed, data and their associated information that have been annotated and grouped could be used to develop phenotype prediction models. In this approach, a machine learning approach can be used to predict the effect of a given compound on a phenotype (described by *in vivo* assay data) that may be toxic or therapeutic or has no effect, using target-based and *in vitro* functional assays as molecular descriptive variables for the model. If relevant data are not available in the ChEMBL database, work could be undertaken to fill data gaps by prediction of protein target inhibitors (e.g. those described in HeCaToS Deliverable D1.5) or prediction of metabolites of compounds (e.g. those described in HeCaToS deliverable D1.6). Ultimately, exposure and/or pharmacokinetic data should be integrated into the graph database and, for instance, this may help to explain why two compounds that inhibit the same toxicity-related protein target(s) do not necessarily lead to the same toxic phenotypes.

## **2. Selection of data from the ChEMBL database**

The graph database is based on a subset of the ChEMBL database. The data focussed on the binding assays, *in vitro* functional assays and *in vivo* assays and the following protocol was used:

1. A set of single protein targets was considered as a starting point and, by extension, all the binding assays in which they have been tested;
2. For each binding assay, each compound that is considered to be active on the protein target is identified;
3. The *in vitro* functional assays are selected if they share at least one of their active compounds with a binding assay;
4. Similarly, the *in vivo* assays are selected if they share at least one of their active compounds with a binding assay.

## 2.1. Selection of assays

### 2.2.1 Binding assay selection

The binding assays represent the least complex type of experiment that determines if a compound interacts with a given protein target and whether this interaction involves inhibition or activation of the protein target. Using a set of 330 protein targets (Annex Table 1), the ChEMBL database was queried to collect every relevant binding assay for a selected set of measured end-points (e.g. percentage of inhibition, IC<sub>50</sub>, K<sub>d</sub>, K<sub>i</sub>, potency, ED<sub>50</sub>, etc.).

Three sets of protein targets were identified according to their influence on the human body:

- Therapeutic targets;
- Cardiotoxic targets
- Hepatotoxic targets

The protein target categories aim to select proteins with known effects when modulated with approved drugs, or with molecular compounds. The selected therapeutic targets correspond to proteins targeted by drugs that are currently on the market. The selected cardiotoxic and hepatotoxic protein targets are known to lead to heart and liver toxicity respectively<sup>2</sup>. The process to identify each category of protein target is given in the Annex.

In addition, a second approach was applied to identify binding assays in which drugs were tested that have been (i) withdrawn from a national market due to hepato- or cardiotoxicity, or because (ii) hepatotoxicity is mentioned in the approved drug labelling<sup>3</sup>. Note that no source of data including cardiotoxicity warnings within the approved drug labelling could be found. In total, 85 unique drugs were found tested in at least one binding assay. Their name with the reason why they have been selected is presented in Annex Table 2.

Because these drugs might be active in binding assays that involve targets that are not part of the initial protein target selection, new proteins were identified that resulted in 686 unique protein targets. The combination of these approaches resulted in the identification of a total of 58,554 unique binding assays (Table 1).

### 2.1.2 *In vitro* functional assays

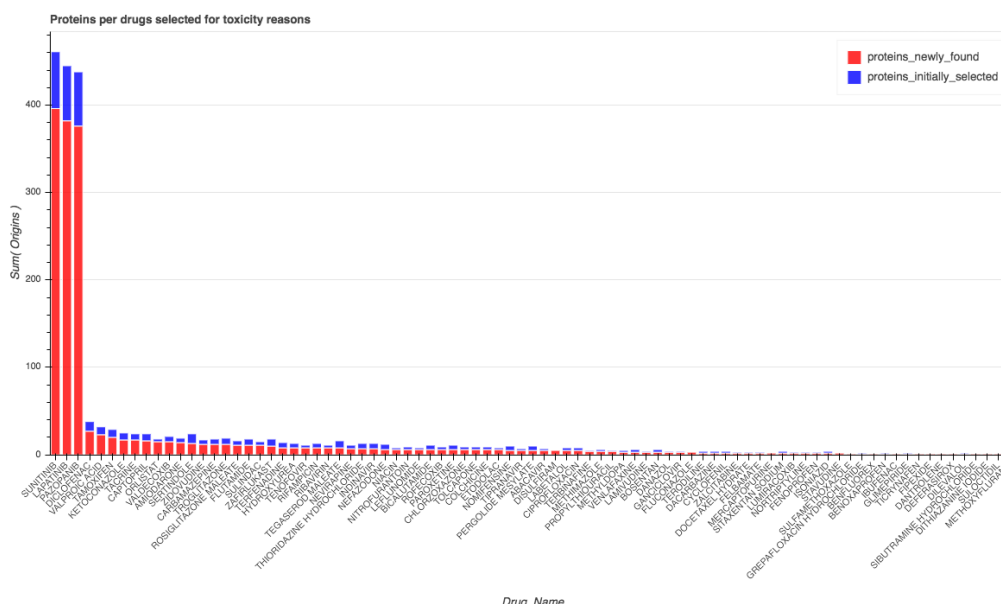
This category contains every assay for which the target is either an isolated cell, a cell-line, an isolated organ or a tissue. Only assays that share the same active compounds were selected as the aim was to see the effect of a given compound on the different assays. Therefore, before selecting any *in vitro* functional assays, the active compounds in the binding assays were first identified (detail explained in Section 2.2). By selecting only mammalian targets and the set of active compounds, 29,367 unique *in vitro* functional assays were identified. In addition, at this stage, care was also taken to filter out all the 'unsuitable' *in vitro* functional assays because they would be related to uninterpretable data as explained in Annex. Hence, a set of 8,157 *in vitro* functional assays was considered.

### 2.1.3 *In vivo* assays

The *in vivo* assays were selected from the ChEMBL database only if they shared at least one compound with a binding assay. The details as to how the binding assay active compounds were identified are presented in the Section 2.2. *In vivo* assays, unlike functional *in vitro* and binding assays, are performed on the living animal, and allow us to observe the effect of a compound on a whole organism. In total, 38,232 *in vivo* assays were identified.

### 2.1.4 Summary of selected assays

The number of assays that were selected for the study is presented in Table 1. This is based on an initial selection of 330 protein targets (Section 2.2.1) as well as the 356 protein targets identified using the toxicity-related drugs (Section 2.2.1). Most of the drugs selected for toxicity reasons were also tested on the protein targets initially selected. However, many other protein targets were identified for some toxicity-related drugs which explains why the number of proteins has more than doubled (Figure 2).



**Figure 2:** Number of proteins found for the drugs selected for toxicity reasons.

The binding assays represent the biggest set of assays with almost 60,000, followed by the *in vivo* assays (more than 38,000) and the functional *in vitro* assays (almost 30,000). Although, the ChEMBL database contains about two times more functional (*in vitro* and *in vivo*) assays than binding assays, the selected assays show a ratio of almost 1:1. This means that the range of active compounds that are tested in binding assays are not systematically tested in functional *in vitro* and *in vivo* assays.

**Table 1:** Summary of the numbers of selected assays after each step. \*The 330 protein targets initially selected and 356 other proteins derived from the withdrawn drugs are included (Section 2.2.1).

Steps	Protein targets*	Binding assays	Functional <i>in vitro</i> assays	<i>In vivo</i> assays
Initial selection	686	58,554	29,367	38,232
After annotation	686	58,554	8,157	20,714
After active compound selection	686	26,771	1,859	20,714

### 2.2. Selection of active compounds

The procedure used to select the three sets of assays was described in the section 2.1. However, *in vitro* functional and *in vivo* assays were selected only if at least one of the compounds found active in a binding assay was also active in one of these assays. The reason for this choice is to understand the effect of an active compound from the molecular level to the organism level. Therefore, considering only the cases in which these compounds are active in all the different types of assays appears as the best solution to reach this objective.

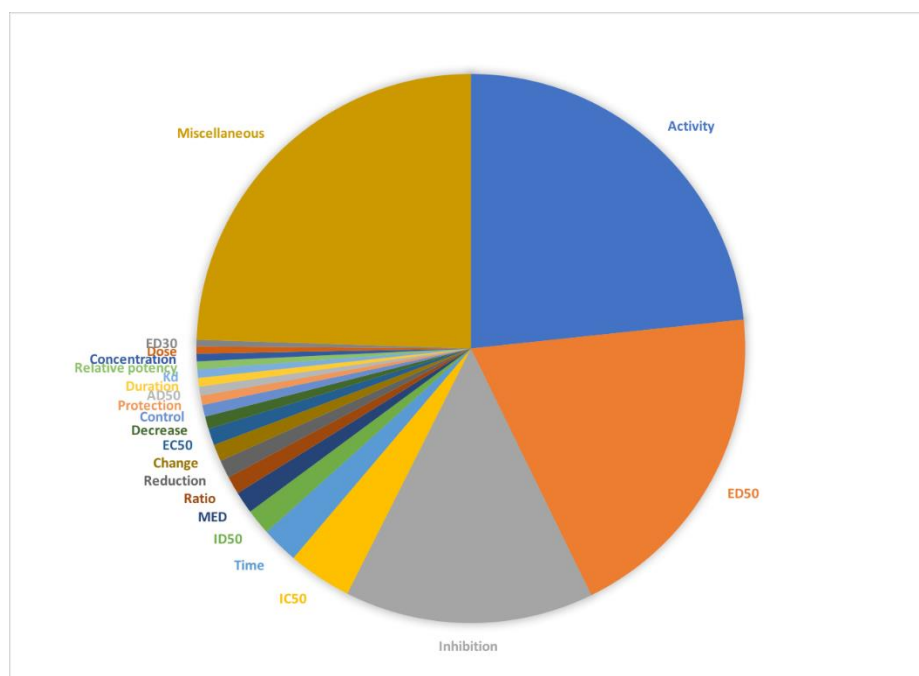
In this report, for the binding assays and the *in vitro* functional assays, an ‘active’ activity threshold was defined using the pChEMBL values. The pChEMBL value<sup>4</sup> is the negative logarithm of the activity

value (in molar units) for dose response activity types such as IC50, XC50, EC50, AC50, Ki, Kd and Potency (with some quality filtering being applied).

For the binding assays, using a threshold of  $\geq 5$  for the pChEMBL values, 131,357 unique parent molecules were identified where a parent molecule is the putative bioactive component<sup>5</sup>.

To check if these compounds were also active in the *in vitro* functional assays, the pChEMBL values were also used, considering a slightly lower threshold (or higher if referring the molar value). It is a well-known observation that a higher concentration is typically required to obtain an effect in a cell based assay than in a protein binding assay. This is explained by the non-specific interactions the compound may have with the different constituents of the cell. Hence, a pChEMBL threshold of  $\geq 4$  was used, and 7,615 active compounds were found.

Finally, the active compounds in the *in vivo* assays were identified. Unlike the two other assay types, there is usually no associated pChEMBL value. The *in vivo* end-points described by the ChEMBL database are numerous, with three different activity types that account for the three quarters of the *in vivo* data (Figure 3). Moreover, the activity types 'Activity' and 'Inhibition' are mainly associated with a percentage unit or no unit. Therefore, they represent the evolution of a function or a behaviour following the introduction of a given dose of compound. As for the binding assays, there are data that it is better not to use because they are considered not accurate enough for modelling. Hence, only the ED50 could have been used if the same criteria as for the binding and *in vitro* functional assays was applied.

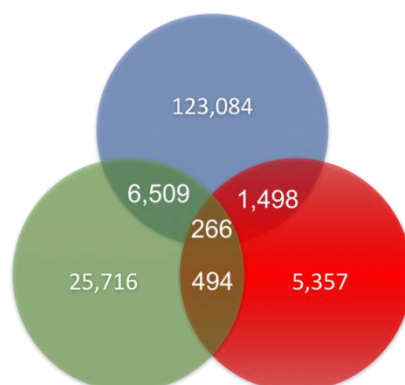


**Figure 3:** Proportion of activity types associated with the selected *in vivo* assays.

Therefore, active molecule selection in the *in vivo* assays was made by considering all activity types associated with a numerical value. This choice is not perfect but was necessary given the complex and diverse end points measured. Every activity with no numerical value but that was annotated as being active by the authors of the study was also included in the selection. That way, 32,985 unique compounds were found.

Comparison of the three sets of active compounds (Figure 4), shows that a large majority of the compounds (162,924) has been detected as active only once among the three types of assay. Shared

active compounds are more abundant between the binding assays and *in vivo* assays (6,509) than between the binding assays and *in vitro* functional assays (1,498). This might be due to the fact that the selection of the actives on the *in vivo* assay actives was less restrictive. However, the low number of shared active compounds between the *in vitro* functional assays and *in vivo* assays (494) seems to indicate that few compounds are either tested or active in both types of assays. The number of active shared by the three sets is also relatively low (266).



**Figure 4:** Venn diagram of the sets of active compounds for the binding assays (blue), the *in vitro* functional assays (red) and the *in vivo* assays (green).

Despite the protocols applied to select the maximum number of assays that have active compounds in common, the majority of the molecules are restricted to only one type of assay. However, there is an aspect that has not been taken into account yet: the molecular similarity.

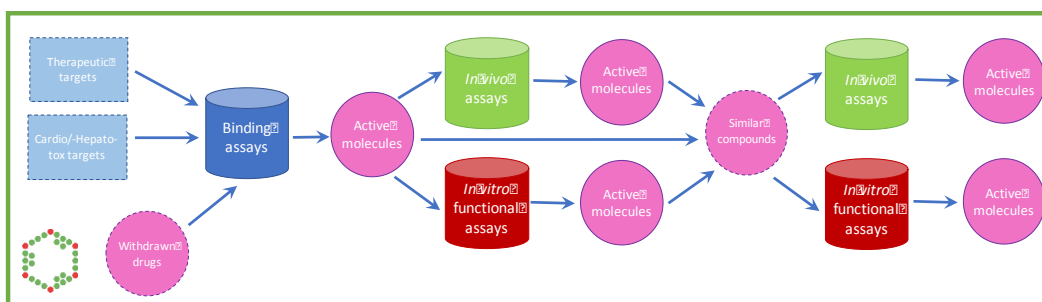
### 2.3. Data selection expansion through molecular similarity

In order to increase the number of shared active molecules between the different sets of assay, and therefore, increase the number of related assays, the molecule similarity of the active compounds was considered. Molecular similarity may be defined and calculated using several different approaches<sup>6</sup>. Here, the expression ‘molecular similarity’ will be used to encompass two distinct methods described in detail in the Annex section:

- Tanimoto coefficient;
- Bemis and Murcko scaffold sharing.

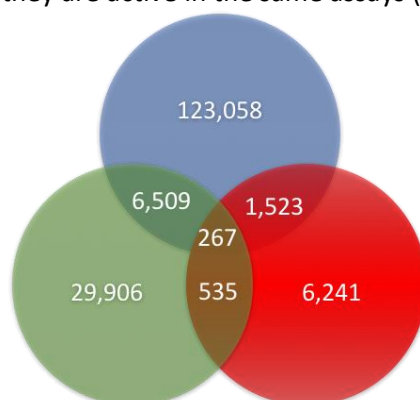
The fingerprints were calculated for the whole set of active molecules and a pairwise Tanimoto similarity calculation was performed. Additionally, to extend the number of selected compounds, all ChEMBL molecules were considered by making a comparison between them and the set of active molecules. Using the molecular scaffold, a similar approach was performed by first making the comparison analysis on the selected active compounds, then by comparing their scaffolds to all the unique scaffolds found in ChEMBL molecules.

In this case, two molecules are considered as being similar if their Tanimoto coefficient is greater than 0.8 and if they share the same scaffold. Additionally, the comparison with the other ChEMBL compounds allows retrieval of a larger proportion of the compound database. From these new compounds identified, a second round of data extraction was undertaken using the same parameters as before. It consisted of the identification of new *in vitro* functional and *in vivo* assays on which these new compounds are active (no new binding assays were found because every possible active of these assays was already included). Finally, all the active molecules of the newly identified assays were selected. The whole process is illustrated in **Figure 5**. The selection of similar molecules could continue over several other iterations but the choice was made to stop because the consequence would be to select compounds that are more and more dissimilar to the active compounds found in the binding assays.

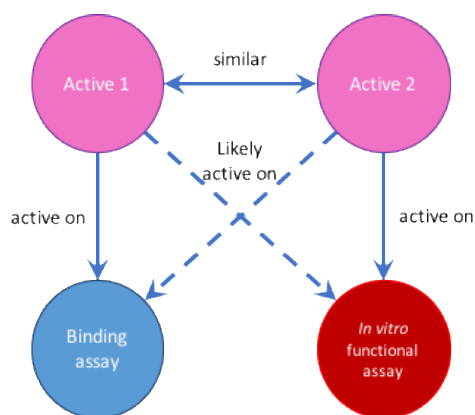


**Figure 5:** Data selection workflow. Binding assays are identified from sets of therapeutic, cardiotoxic and hepatotoxic protein targets, and from sets of withdrawn drugs due to cardiotoxicity or hepatotoxicity. Active compounds tested in the binding assays are used to select sets of *in vitro* functional and *in vivo* assays. All other active compounds tested in these assays were also selected. The first iteration identified similar compounds from the ChEMBL database and the second iteration identified other sets of *in vitro* functional and *in vivo* assays in which these compounds are active.

A total of 168,840 unique compounds (163,722 before the similarity-based expansion) were selected. Looking at the same compounds shared between the assay sets (Figure 6) and comparing with that obtained previously (Figure 4), it was observed that the overlaps did not increase much. Between the binding assays and the *in vitro* functional assays 25 additional compounds were found, and 41 additional compounds between *in vitro* functional and *in vivo* assays. However, the key improvement is that the similarity information that was collected can be used to link assays that do not share the same active compound but whose compound similarity is considered similar enough to be sufficiently confident that they are active in the same assays (Figure 7).



**Figure 6:** Venn diagram of the sets of active compounds after similarity-based expansion for the binding assays (blue), the *in vitro* functional assays (red) and the *in vivo* assays (green).



**Figure 7:** Illustration showing that two similar compounds that are independently active on a binding assay and an *in vitro* functional assay respectively, may also be used to form connections between these assays that would not otherwise be observed.

### 3. Annotation of selected data and its representation as a graph database

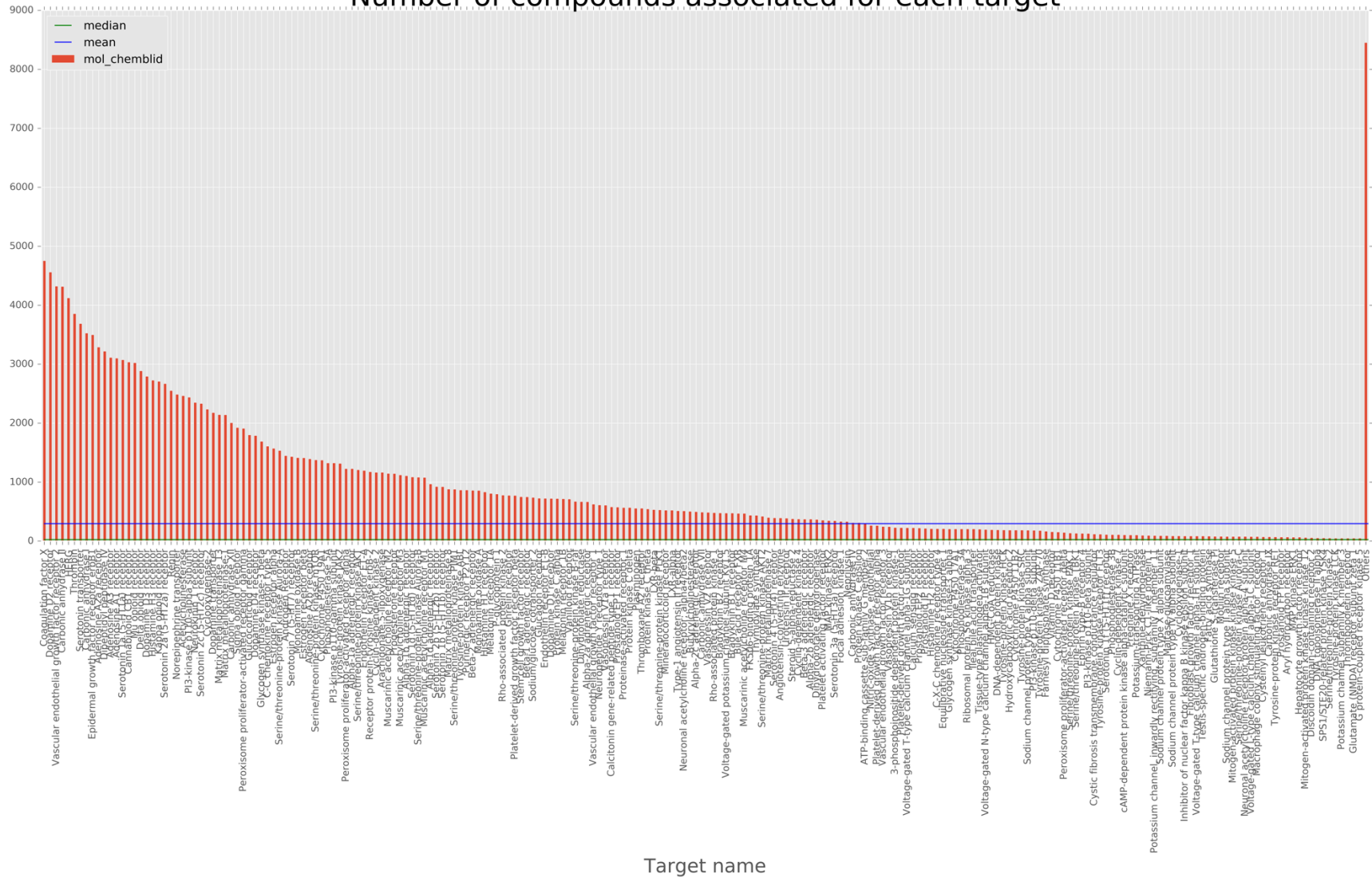
Navigation through the large amount of data contained in the ChEMBL database is not straightforward even if only a small fraction of the data is considered. Instead, the approach taken has been to group data that describes a common entity such as a similar protein target, or a similar *in vivo* assay. Note that the *in vitro* functional assays have not yet been annotated in the ChEMBL database, as discussed in Section 5.

#### 3.1. Annotation of the protein targets

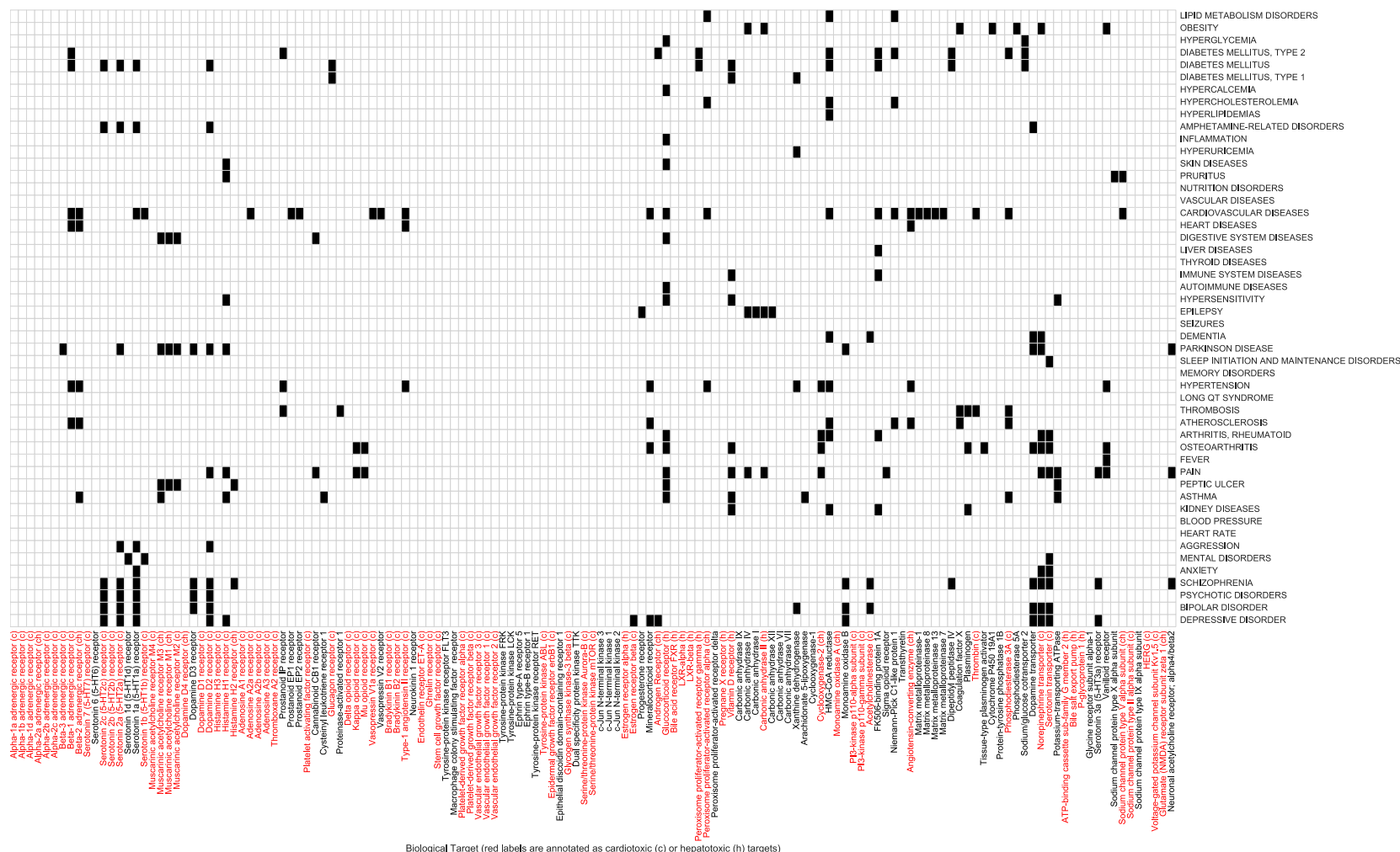
Binding assays describe the interaction of a compound with a protein target. Therefore, the compounds associated with each binding assay can be reorganised and grouped by protein target. Using this approach, one or more binding assays can be linked to a single protein target and therefore the active compounds found for an assay can be associated with their corresponding protein target. As a result, 686 distinct protein targets that are associated with one or more active compounds were identified (Figure 8). Note that the mean number of active compounds per protein target is 297, but that the median is 25 and therefore there are many compounds associated with a small number of protein targets. Using the disease names associated with each protein in the ChEMBL database, the proteins were then annotated (see Annex for full details).

Therefore, the protein targets can be directly mapped to the indications of the drugs that interact with them. An example heatmap is provided in Figure 9 for a set of protein targets and corresponding drug indications described by their MeSH terms (MeSH version 2017; <https://www.nlm.nih.gov/mesh/>), where each black cell represents one or more drugs that link a protein target to an indication.

## Number of compounds associated for each target



**Figure 8:** Bar chart of the number of active compounds associated with each protein target. The 467 proteins with less 50 compounds per target have been grouped together and are labeled 'Others'. The mean (blue line) and the median (green line) have been calculated on the overall proteins.



**Figure 9:** Heatmap of drugs that link between the approved drug indication (MeSH term) and its protein target. An example is given for the 143 protein targets and 50 drug indications that have also been used to annotate the in vivo MeSH phenotypes (Section 4.4). A black cell indicates that one or more marketed drugs provide a therapeutic link between a protein target and its drug indication.

### 3.2. Annotation of *in vivo* assays

The *in vivo* data that are described in the ChEMBL database include a wide range of assays used to investigate differing phenotypes. For example, different *in vivo* assays may investigate:

- Anesthesia or pain using the 'tail-flick' assay in mouse ('AD50' activity type; mg kg<sup>-1</sup>);
- Anticonvulsant activity in mouse to test for an effective dose ('ED50' activity type; of mg kg<sup>-1</sup>);
- Anti-inflammatory activity in rat using the 'paw edema' test ('inhibition' activity type; %), or
- Test for hepatic damage in rat ('protection' activity type; %).

Given that there are around 38,000 selected *in vivo* assays, it is desirable to group similar *in vivo* assays so that the data relating to one phenotype may be placed into a similar category. For example, if all *in vivo* assays that investigate 'inflammation' are tagged, or annotated, with the same category name then these assays and their associated activities can be grouped together for further analysis.

Annotation of the selected *in vivo* assays was done using a comprehensive reference work called 'Drug Discovery and Evaluation: Pharmacological Assays'<sup>7,8</sup> (edited most recently by Hock in 2016) which describes most functional assays in substantial detail. Hock<sup>8</sup> describes about 1000 assays that may be classed as functional *in vitro*, *ex vivo* or *in vivo* assays. For each assay, Hock provides an assay name, purpose and rationale, procedure, evaluation, critical assessment of the method, modifications of the method, references and further reading. In addition, similar assays are organised by chapter. For example, the chapter on "Cardiovascular Analysis *In Vivo*" contains assays that investigate blood pressure by different methods, angiotensin II antagonism or the Bezold-Jarisch reflex that causes excessively shallow breathing or an abnormally low resting heart rate.

The description of each *in vivo* assay given by Hock has been used as a basis to group the *in vivo* data selected from the ChEMBL database. There is no standard wording for the description of each *in vivo* assay, and these assay descriptions vary in vocabulary, syntax, length and usefulness, depending on the information available in the primary literature source. Therefore, the approach taken has been to find a text pattern that uniquely identifies one 'reference' *in vivo* assay described by Hock (2016) and to match this pattern against the text description contained within the 'assay\_description' field of the selected *in vivo* assays. For example, the regular-expression text pattern 'Tail\W?Flick' identifies the reference 'Tail Flick' assay described by Hock (2016), and allows the annotation of all *in vivo* assays that have a relevant 'assay\_description' e.g. "Analgesic activity in tail flick test, oral administration", or "Compound was administered subcutaneously and was evaluated for opioid antagonist activity (versus morphine) by tail-flick (TF) antagonism test". Examples of selected *in vivo* assays that have been annotated using text patterns that annotate at least one reference *in vivo* assay described by Hock are presented in Figure 10.

	annotated_assay_pattern	assay_chembl_id	assay_description	target_organism
annotated_assay_name				
Blood Glucose Lowering Effect, General Antidiabetic Activity	Blood\W?Glucose\Plasma\W?Glucose, Diabet	CHEMBL1121054	Antidiabetic activity in Wistar albino rat assessed as reduction of blood glucose level administered qd for 9 days by glucose oxidation method (RVb = 79.17 +/- 0.48 mg/dl)	Rattus norvegicus
Bone Anabolic Activity in Ovariectomized, Osteopenic Rats, General Hormonal Activity	anabolic.*bone, anabol	CHEMBL1064192	Osteoanabolic activity in ovariectomized Sprague-Dawley rat assessed as bone formation rate using calcein at 30 mg/kg, sc qd for 24 days relative to dihydrotestosterone	Rattus norvegicus
Catalepsy in Rodents , General Antipsychotic and Neuroleptic Activity	Catalep\Wring\W.*immobil\immobil.*\Wring\W\Wtetrad\W, Psychot\Neurolep	CHEMBL779609	In vivo antipsychotic activity measured by the induction of catalepsy in rats after po administration.	Rattus norvegicus
Cholesterol Diet Induced Atherosclerosis, General Cholesterol Activity	cholesterol\W?diet\hypercholesterol\hypocholesterol, cholester\WLDL\W\WLDL\W	CHEMBL779174	Percent change in serum HDL level was determined in hypercholesterolemic rats at 10 ug/Kg upon peroral administration	Rattus norvegicus
Compulsive Gnawing in Mice	Gnawing	CHEMBL719262	Tested for stereotyped gnawing and licking behavior of mice pretreated with apomorphine (AG)	Mus musculus
Coronary Artery Ligation, Reperfusion Arrhythmia and Infarct Size in Rats, General Anti-Arrhythmic Activity	Coronary\W?Artery\W?Ligation, Arrhythm\ Fibrillation\ tachycard\ bradycard	CHEMBL778110	Compound that antagonized arrhythmias evoked by coronary artery ligation in the rat was measured at 0.05 mg/kg dose	Rattus norvegicus
Delayed Type Hypersensitivity, General Immunosuppressive Activity	Delayed\W?Type\W?Hypersensit, Immunosuppress	CHEMBL3362879	Immunosuppressive activity in C57BL/6J mouse KLH-induced delayed-type hypersensitivity model assessed as alleviation of footpad swelling at 16 nmol administered on day 7 and 8 post-sensitization measured on day 9	Mus musculus
Diuretic Activity in Rats Lipschitz Test, Saluretic Activity in Rats	Diuretic\ Lipschitz, Saluret\ Kaliuret\ Natriuret	CHEMBL779193	Diuretic effect was evaluated in the conscious female rat by natriuretic assay involving metabolic caging at dose 30 mg/kg	Rattus norvegicus

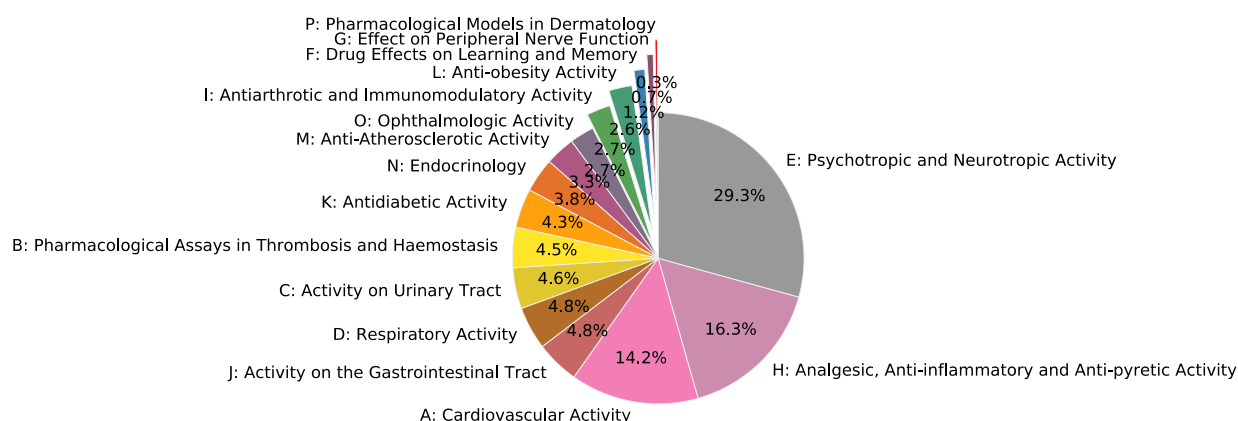
**Figure 10:** Examples of selected, annotated *in vivo* assays. The annotation (in white) gives the annotated\_assay\_name and annotated\_assay\_pattern, while the *in vivo* assays (in yellow) are described by their ChEMBL\_ID, assay\_description and target organism.

The aim of the text pattern matching approach is to annotate selected *in vivo* assays to conform with the ChEMBL database philosophy that maintains high-quality data that accurately reflects its data source. Therefore, for each reference assay described by Hock, a text pattern has been chosen to select the most accurate set of selected, matched *in vivo* assays. Hence, the set of matched *in vivo* assays that correspond to each text pattern has been manually inspected to check that the selected assay results correctly represent the intended Hock reference assay (a positive match) and that the selected assay results do not miss significant numbers of assays that should be annotated by the specified Hock reference assay (a negative match).

Of the 561 reference *in vivo* assays described by Hock that have been used in this work, around half (i.e. 293 reference assays) have a text pattern that allows annotation of selected *in vivo* assays. As a result, 23,679 *in vivo* assays out of a total of 38,232 (62%) were annotated. Further detail is given in the Annex.

The annotated, selected *in vivo* assays have been grouped by phenotypic area based on the chapter of Hock in which the reference assay is described (Figure 11). This shows that the majority of the annotated, selected *in vivo* assays investigate psychotropic and neurotropic activity (29 %), or analgesic, anti-inflammatory and anti-pyretic activity (16 %), with smaller numbers of *in vivo* assays that consider, for example, cardiovascular activity (14 %). This breakdown is considered to reflect the types of phenotypes that lend themselves to *in vivo* investigation and are described in the ChEMBL database.

Annotated *in vivo* assays  
(23697 out of 38232 total assays; ie 62.0 % )



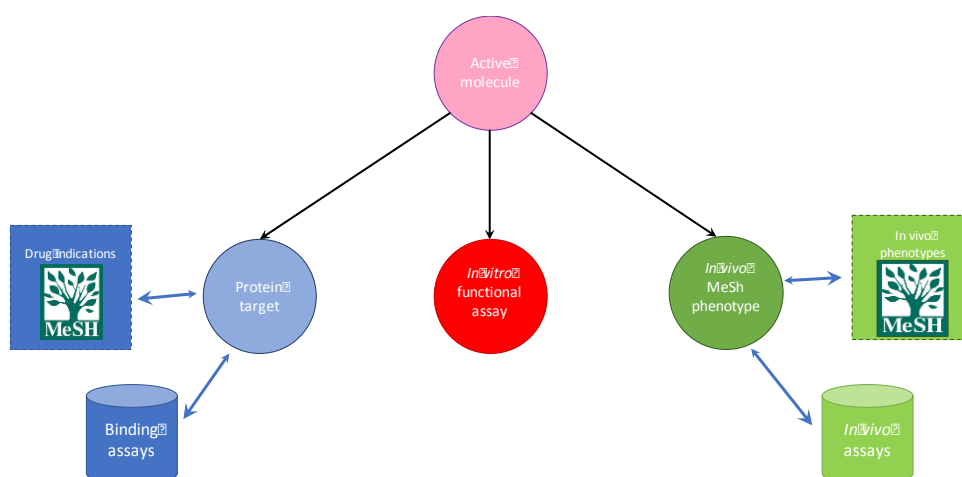
**Figure 11:** Annotated *in vivo* assays selected from the ChEMBL database. There are the 23697 annotated *in vivo* assays (62 %) out of a total of 38232 *in vivo* assays.

A mapping between each selected *in vivo* reference assay given by Hock, and the MeSH term that describes the equivalent phenotype or disease has been carried out. This additional layer of annotation of the selected *in vivo* assays provides the link between the phenotypic outcome observed and the underlying bioassay. The curation was carried out manually using the reference assay description provided by Hock assuming that it was sufficiently detailed to allow a match with a phenotypic disease-related MeSH term. Note that not all *in vivo* reference assays could be annotated by MeSH terms, for example e.g. “Urinalysis”, was not specific enough to be annotated by a MeSH term. Hence, of the 293 reference *in vivo* assays described by Hock, 218 (74 %) also have a MeSH phenotype annotation. The set of *in vivo* reference assays described by Hock annotated with MeSH terms will be referred as the *in vivo* MeSH phenotypes. As a result, the dataset contains 20,714 *in vivo* MeSH phenotypes.

### 3.3. Representation of the selected, annotated data as a graph database

The selected data have been annotated and pooled in order to navigate through them more efficiently. The objective is now to find the relationships that exist between the sets of binding, *in vitro* functional and *in vivo* assays. The expression ‘assay cascade’ is a familiar concept from drug-discovery and it defines the compound progression through assays of increasing complexity and disease relevance. By analogy with this concept, an assay cascade was considered as a protein target, an *in vitro* functional assay and an *in vivo* MeSH phenotype that are linked together by the same active compound (Figure 12). To maximise the number of cascades, partial assay cascades that comprise a protein target and an *in vivo* MeSH phenotype, or an *in vitro* functional assay and an *in vivo* MeSH phenotype were also considered.

The active compounds shared by all the assays provide a means to connect the whole dataset. This approach has been implemented as a graph database in which all the different entities have been represented as nodes and the nodes linked together if they share a property. For instance, compounds can be linked to proteins or assays in which they are active, or binding assays can be associated with their corresponding protein target. The graph database uses the Neo4j software because it offers a free community version, is relatively easy to learn and has good performance<sup>9</sup>. Moreover, Neo4j has an increasing user community focusing on different data-related domains providing helpful information. Querying the graph is possible using the Cypher language that has been developed by Neo4j.



**Figure 12:** Example of assay cascade formed by a protein target (blue), an *in vitro* functional assay (red) and an *in vivo* MeSH phenotype (green), linked together by a compound active in each of them (pink). The protein target is linked to a set of binding assays in which it has been tested. Similarly, the *in vivo* MeSH phenotype is linked to a set of *in vivo* assays. Both the protein target and *in vivo* MeSH phenotype are annotated with MeSH terms

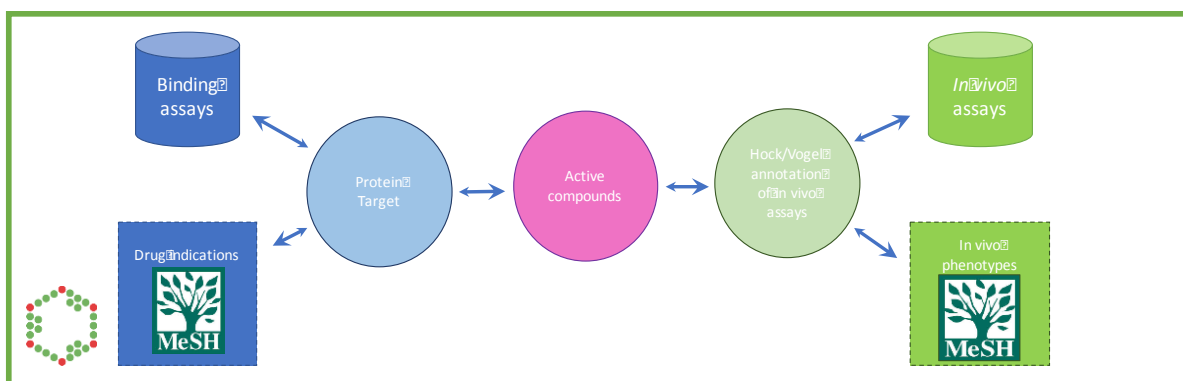
In the graph representation of the selected data, protein targets, binding assays, *in vitro* functional assays, *in vivo* MeSH phenotypes, *in vivo* assays, the active compounds and the MeSH annotations represent the nodes, with each node being labelled to identify which data it belongs to (Figure 12). Additionally, a node has an identifier to allow easy identification (*i.e.*: to distinguish the protein targets, the active compounds or the *in vivo* MeSH phenotypes). Edges of the graph link the active compounds to the protein targets, *in vitro* functional assays, annotated *in vivo* assays only if the compound is active on the corresponding data (Figure 12). Each protein is associated with a set of binding assays, according to which target was tested in the assays, and a set of MeSH annotated derived from the approved drug indications. Similarly, the *in vivo* MeSH phenotypes are linked to the *in vivo* assays that they represent and to their corresponding MeSH terms.

Because the data has been derived from different sets of protein targets and also from drugs with serious toxic effects on the liver and heart, this information can be used to find the related assay cascades. Drugs of interest for the HeCaToS project (Annex Table 3) can also be used to navigate through the graph, and this is described in Section 4.1.

### 3.4. Linking a protein target to its *in vivo* phenotypic outcome

The graph representation of the selected data described in Section 3.3 has been used to investigate the relationship(s) between protein targets and their *in vivo* phenotypic outcomes. It has been assumed that the interaction between an active compound and a protein target will manifest itself as a relevant *in vivo* phenotypic outcome if data from a whole animal experiment is available. Therefore, it should be possible to use the data described in the ChEMBL database to observe a correlation between active compounds that interact with a protein target and active compounds that result in an observed *in vivo* phenotypic outcome.

Using the graph representation, the active compounds for the set of binding assays that represent each protein target have been collated (see the top left, blue cylinder in Figure 13). Equally, the compounds for the set of *in vivo* assays that represent each *in vivo* MeSH phenotype have been collated (see top right, green cylinder in Figure 13).



**Figure 13:** Using the graph database representation of the selected data to link between protein targets and their *in vivo* phenotypic outcomes.

Note that the set of *in vivo* assays contains all compounds (not only active compounds) for each *in vivo* phenotypic endpoint since a threshold activity cut-off is difficult to define. For the data considered in this section, this gives 611 protein targets that are linked to 88 *in vivo* MeSH phenotypic endpoints by 6509 shared compounds.

A Spearman rank correlation has been calculated for compounds that are common between the set of compounds for a protein target, and the set of compounds for an *in vivo* MeSH phenotypic endpoint. For example, if there are 9 compounds that have activities for a specified protein target (e.g. cyclooxygenase-2), and of these, 5 compounds also have activities for a specified *in vivo* MeSH phenotypic endpoint (e.g. 'pain', with an activity type of ED50 in mg kg<sup>-1</sup>), then the activities of the 5 shared compounds are used to perform a Spearman rank correlation. The Spearman rank correlation coefficient was carried out if there were 4 or more datapoints. Note that the Spearman rank correlation could be calculated for 3 shared compounds per biological target to *in vivo* MeSH phenotype pair, but this gives overall results that are more likely to be strongly positively or negatively correlated because of their small sample size, leading to data that is visually skewed in the final plot.

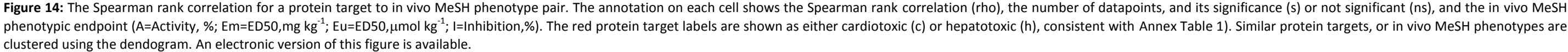
It is difficult to set a threshold for active compounds for an *in vivo* MeSH phenotype because there are multiple activity types and unit combinations for *in vivo* assays, which would require a different activity threshold to distinguish active from inactive compounds. The four main *in vivo* MeSH phenotypes described by the ChEMBL database are:

- Activity (%);
- ED50 (mg kg<sup>-1</sup>);
- ED50 (μmol kg<sup>-1</sup>);
- Inhibition (%);

Numerous other activity types and units apply to small numbers of assays (e.g. 'TIME', 'MED', 'ID50', 'REDUCTION', see Figure 3).

As a result, the approach was taken to assume that all compounds that annotated as *in vivo* MeSH phenotypic endpoints were active if an activity value was present for the activity types of 'Activity' or 'Inhibition'. For an ED50 activity type, all activity values were included unless they had a more than sign ('>'), in which case they were excluded due to their low potency. For each protein target to *in vivo* MeSH phenotype pair, the most significant Spearman rank correlation result (i.e. smallest p value) was chosen from the four possible *in vivo* MeSH phenotypic endpoints, to be included within an overall heatmap plot. Clustering of similar protein targets, or *in vivo* MeSH phenotypes, was performed and has been visually shown as a dendrogram on the heatmap. The clustering process is described in the Annex.

The overall Spearman rank correlation for each protein target to *in vivo* MeSH phenotype pair is displayed in Figure 14 (with an accompanying pdf version provided), and is discussed in Section 4.4.



## 4. Results

### 4.1. Using the graph database to investigate HeCaToS drugs

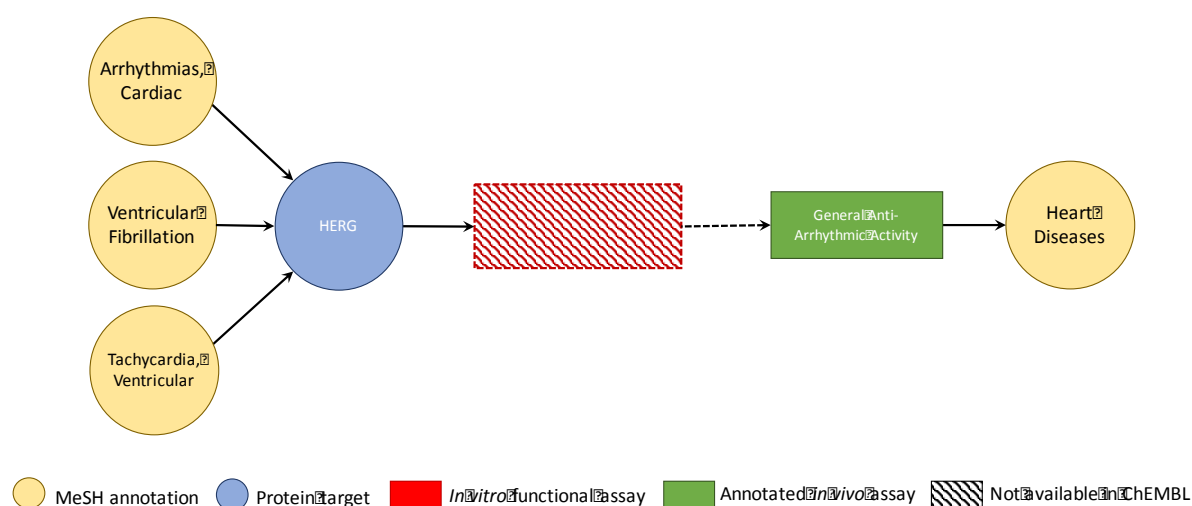
For each of 32 HeCaToS compounds (Annex Table 3), the graph database was searched for an assay cascade to see if the data selection was sufficient to retrieve the therapeutic effect of these drugs. Any known side effects that could be attributed to these compounds were also investigated. The graph was queried for each of the HeCaToS compounds, to check if the resulting assay cascade was coherent. Ideally, each assay cascade should consist of three elements:

- A protein target;
- An *in vitro* functional assay;
- An *in vivo* MeSH phenotype.

The partial assay cascades in which the *in vitro* functional assay or the target was missing were also considered. Although the full assay cascades were preferred, the information provided by these partial cascades is also of interest. On the 32 drugs selected for the HeCaToS project, eight of them were found to have at least one corresponding assay cascade. Each of these cascades is described below.

**Amiodarone:** Only used to treat life-threatening heart rhythm disorders, amiodarone is a hERG blocker and it helps reducing ventricular tachycardia or ventricular fibrillation. Hence, it is not surprising to find an assay cascade showing this drug active on hERG and also in a *in vivo* model of anti-arrhythmia (Figure 15). No *in vitro* functional assay could be found to link the molecular mechanism to the observable phenotype, but the annotations are sufficiently clear to be confident of the relationship that exists between the protein target and the *in vivo* assay. Amiodarone may lead to cardiotoxicity by causing bradycardia<sup>10</sup> but no related data was found in the selected data. Note that amiodarone also has other side effect including pulmonary toxicity<sup>11</sup>.

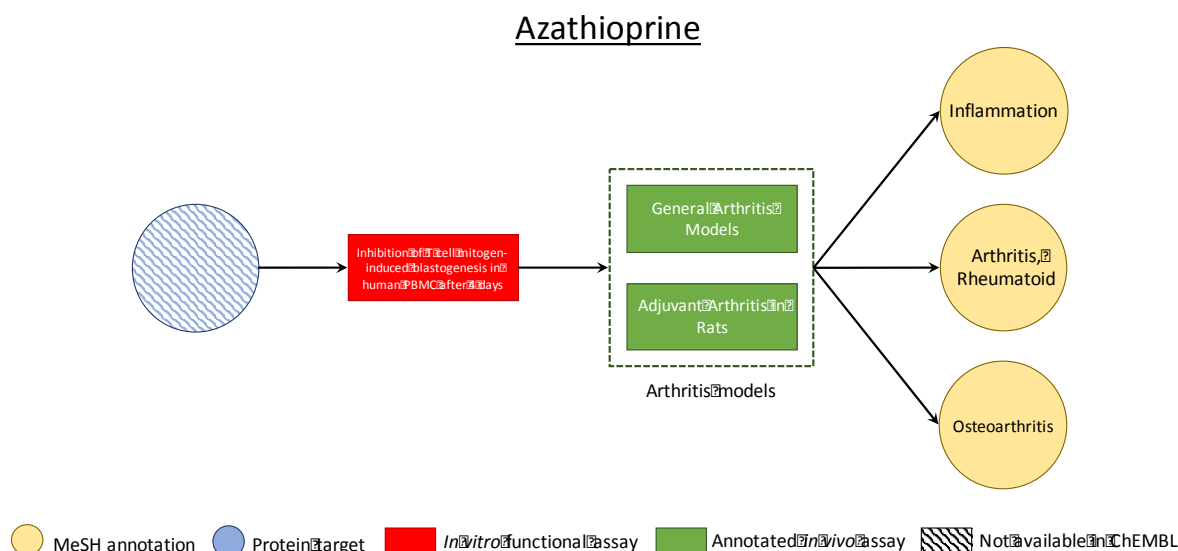
#### Amiodarone



**Figure 15:** Assay cascade related to amiodarone. No activity of this drug on an *in vitro* functional assay present in the graph database was found.

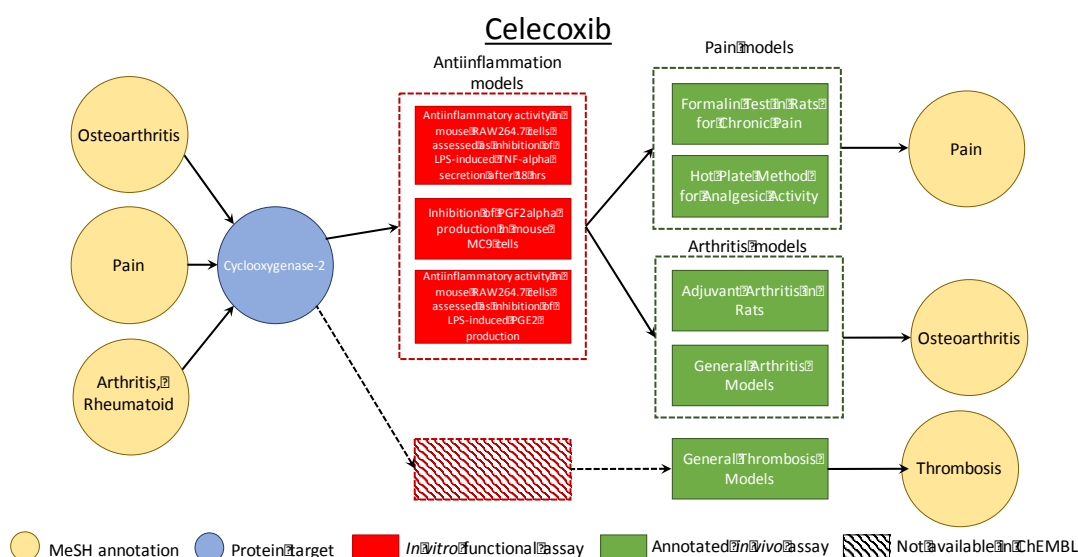
**Azathioprine:** A partial assay cascade showing the effect of this drug on rheumatoid arthritis was found (Figure 16). This drug is used to prevent kidney transplant rejection but also to reduce signs of rheumatoid arthritis. It antagonizes purine metabolism and may inhibit synthesis of DNA, RNA, and proteins, leading to cell death. The identified assay cascade is related to the latter. Indeed, it shows the effect of the compound on T cell and on *in vivo* models of arthritis. Although the protein target information is missing, the information provided by the assays are sufficiently detailed to be

reasonably confident of this cascade. No information was found that could demonstrate the hepatotoxicity of azathioprine.



**Figure 16:** Assay cascade of azathioprine. No activity of this drug on a protein target present in the graph database was found.

**Celecoxib:** This drug is used to treat pain and inflammation caused by diverse conditions such as osteoarthritis or menstruation. The mechanism of action of celecoxib is believed to be due to the inhibition of prostaglandin synthesis and the retrieved cascade illustrates this effect (Figure 17). The inhibition of the protein target Cyclooxygenase-2 leads to an inhibition of factors related to inflammation such as TNF-alpha or prostaglandins. The direct effect is the reduction of the inflammatory response that can be observed in osteoarthritis and certain pains, a phenomenon observed *in vivo*. Interestingly a second cascade was found that may confirm the cardiotoxic property of this drug. Indeed, celecoxib may be at the origin of thrombosis in *in vivo* assays. A report from 2002 has detailed how, by inhibition of Cyclooxygenase-2, the drug induces an increase of thromboxane leading to vascular and thrombotic events<sup>12</sup>. Despite the absence of an *in vitro* functional assay that would show an increase of thromboxane after addition of celecoxib, an existing link between Cyclooxygenase-2 and thrombosis might be suggested.



**Figure 17:** Assay cascades of celecoxib. No activity of this drug on an *in vitro* functional assay was found in the graph database to explain the relationship between the inhibition of the cyclooxygenase-2 and thrombosis.

**Cyclosporine:** Cyclosporine is an immunosuppressant used to prevent organ rejection in people who have received a transplant. Its mode of action consists of the binding of the cyclophilin that leads to the inhibition of the phosphatase activity of calcineurin which in turn is required for the activation of transcription factors that up regulate the expression of inflammatory cytokines. The assay cascades that were found provides insight as to how this drug acts (Figure 18). Cyclosporine inhibits the protein target Cyclophilin A and this interaction reduces the production of pro-inflammatory molecules such as interleukin-2 (IL-2), a link already reported<sup>13</sup>. *In vivo*, it is observed that the drug induces a reduction of the immune system response. No further information was found that could explain the hepatotoxic effect of this drug.

## Cyclosporine

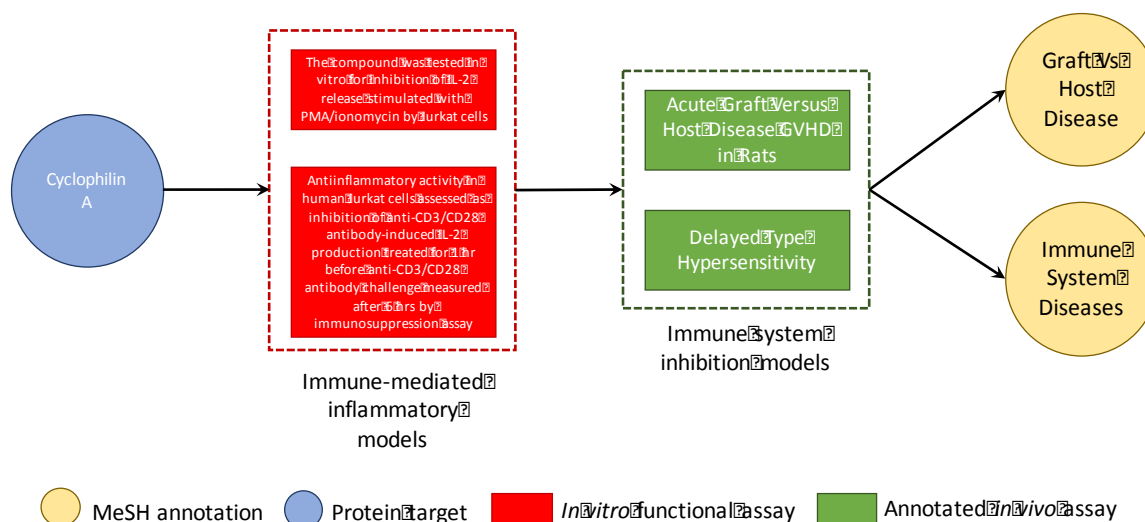


Figure 18: Assay cascade of cyclosporine.

**Diclofenac:** Diclofenac is a non-selective cyclooxygenase inhibitor. Its action helps reduce pain and inflammation. It is used to treat pain but also symptoms of osteoarthritis and rheumatoid arthritis. The graph database allows retrieval of the cascades that correspond to its mode of action, although, related in vitro functional assays are missing (Figure 19). The drug inhibits Cyclooxygenase-2 and also shows an effect on in vivo assays in which pain or arthritis were induced in animal models. The LiverTox database notes that there are multiple factors that cause the hepatotoxicity of diclofenac ([www.LiverTox.nih.gov](http://www.LiverTox.nih.gov)).

## Diclofenac

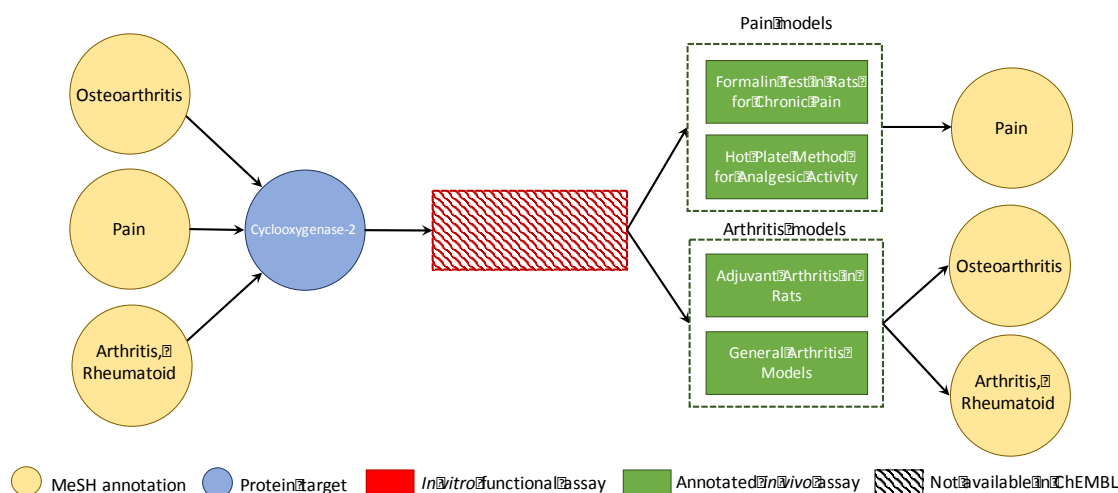
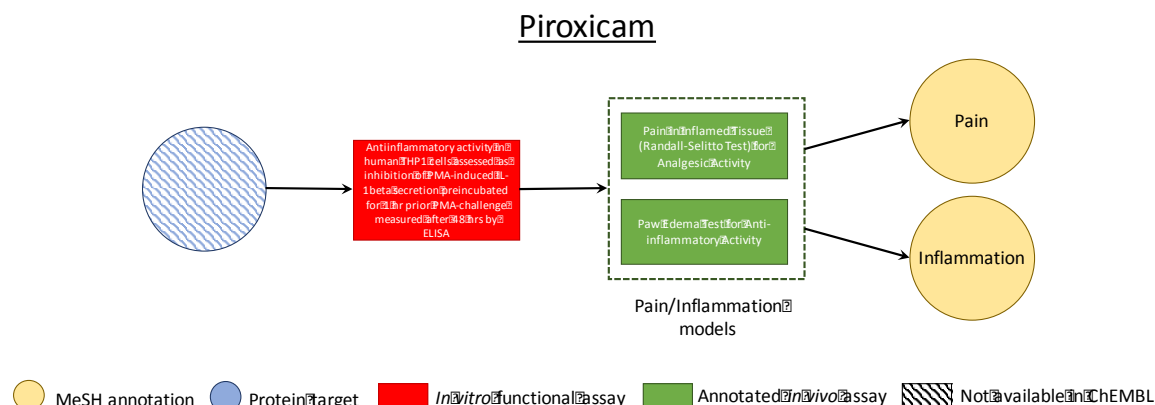


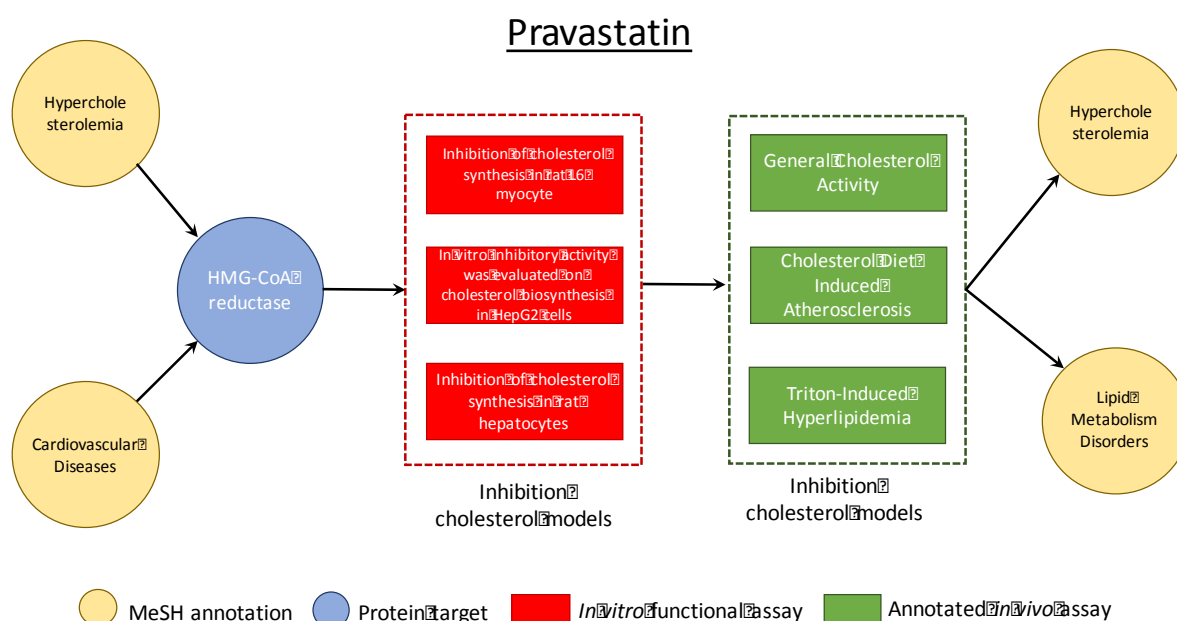
Figure 19: Assay cascade of diclofenac. No activity of this drug on an in vitro functional assay was observed in the graph database.

**Piroxicam:** Piroxicam is another NSAID used to treat pain and inflammation that is likely to act on prostaglandin synthesis leading to a diminution of inflammatory molecules. The observed assay cascade depicts the process (Figure 20). *In vitro*, piroxicam acts by inhibiting the secretion of interleukin-1beta leading to a decrease of the inflammatory response. This effect is confirmed in *in vivo* assays where the compound reduces both pain and inflammation in animal models. Unfortunately, the data selection did not allow identification of the protein target with which the drug interacts. However, it is known to interact with cyclooxygenase-1 and -2<sup>14</sup>. No data related to the hepatotoxic effect of this drug could be found



**Figure 20:** Assay cascade of piroxicam. No activity of this drug on a protein target present in the graph database was observed.

**Pravastatin:** Part of the statin drug family, pravastatin is a substituent of the endogenous substrate of HMG-CoA reductase. The inhibition of this enzyme leads to a diminution of the cholesterol and triglycerides in the blood. The observed assay cascade is consistent with the mechanism of action of this drug (Figure 21). It acts by inhibiting the protein target HMG-CoA reductase resulting in a lower concentration of cholesterol *in cellulo*. *In vivo* experiments show it also leads to less circulating cholesterol. No evidence of hepatotoxicity was found in the graph.



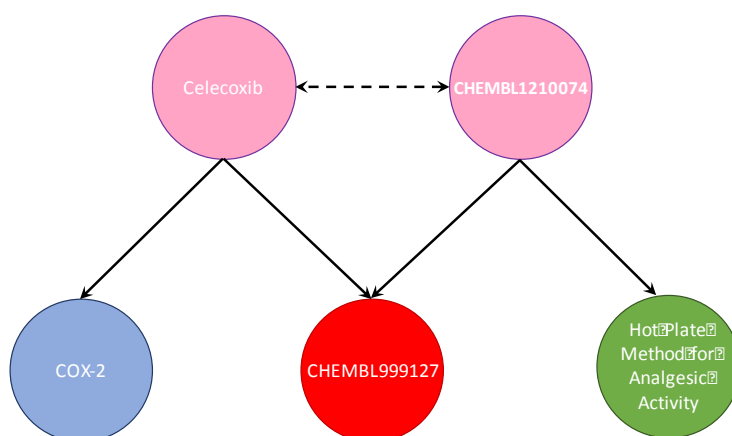
**Figure 21:** Assay cascade of pravastatin.

**Simvastatin:** The corresponding assay cascade is similar to the one found for pravastatin and therefore is not shown here. Both drugs belong to the class of statins and their mode of action is similar. Similar to pravastatin no information relating to the hepatotoxicity of this drug in the graph database could be found.

#### 4.2. Similarity based cascades

Section 2.3 described the addition of similar compounds with the objective of increasing the number of relationships between the assays. Therefore, the similarity information should help to retrieve more assay cascades for the 32 HeCaToS drugs. Using the same approach as in section 4.1, cascades were considered where different compounds are active on the different assays as long as the compounds are similar (Figure 22).

The results showed it was possible to find other assay cascades for the eight HeCaToS drugs in which similar compounds are active on the same targets or the same assays. However, these new cascades did not provide more insight to explain the toxic effects of these drugs. Moreover, no cascade was found for the 24 remaining HeCaToS drugs.



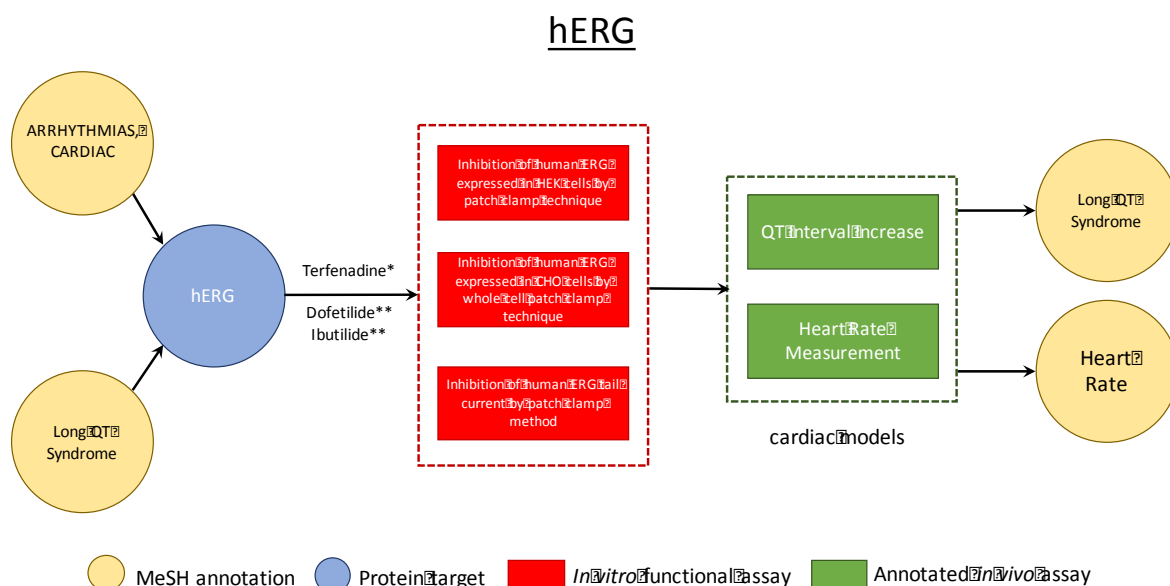
**Figure 22:** Example of cascades obtained using active compound similarity. Each active compound taken individually only forms a subcascade consisting of two components. The association of the two subcascades creates a full cascade of three components. A solid line symbolises that the compound is active, a dashed-line means the compounds are similar.

#### 4.3. Known toxicity mechanisms

The graph database can also be used directly to search for toxicity mechanisms. As a proof of concept, the assay cascades that describe the cardiotoxic effect of the protein hERG were investigated. To do this the graph database was queried to look for every assay cascade with the hERG protein as the protein target and where the *in vivo* MeSH phenotype is “QT interval increase” or “Heart rate measurement”.

As a result, the same cascade for three different compounds was obtained: terfenadine, dofetilide and ibutilide (Figure 23). All three compounds inhibit hERG and this has the effect to perturb the potassium current as confirmed by several *in vitro* functional assays. Moreover, these three compounds interfere with heart rate and QT interval *in vivo*. This illustrates the mechanism of action of hERG inhibitors. Interestingly, dofetilide and ibutilide appear to be both antiarrhythmic agents acting as hERG blockers while terfenadine is an antihistamine targeting H1-receptor. Therefore, while the aforementioned cascade represents the therapeutic mechanism of action of dofetilide and ibutilide, it depicts one of the side effects associated with terfenadine.

In summary, although the graph database does not discriminate between the therapeutic effect and the side effect of a drug, it can be used to find assay cascades involved in toxic effects. The example of hERG is a proof of concept and similar studies for different toxicity mechanisms are possible but not presented here.



**Figure 23:** Assay cascade related to hERG inhibition. \* the cascade corresponds to the side effect of the indicated drug; \*\* the cascade corresponds to the mode of action of the indicated drug.

#### 4.4. Linking a protein target to its *in vivo* phenotypic outcome using a Spearman rank correlation between common compounds

The Spearman rank correlation for each protein target to *in vivo* MeSH phenotype pair is displayed in Figure 14, and shows that the approach can identify relevant therapeutic signals from the selected ChEMBL data applied in this work. For example, Figure 14 shows:

- Medium to strong Spearman rank correlation between protein targets such as serotonin receptors, dopamine receptors, adrenergic receptors and muscarinic receptors that are known to be involved in the brain e.g. [15, 16, 17, 18], and *in vivo* MeSH phenotypes that describe brain disorders such as mental disorders, anxiety, schizophrenia, psychotic disorders, bipolar disorder and depressive disorder (see bottom left corner of Figure 14). Note that dopamine receptors, muscarinic acetylcholine receptors, and serotonin receptors have marketed drugs (phase 4) for many of these *in vivo* MeSH phenotypes (Figure 9)
- Medium to strong Spearman rank correlation between protein targets such as serotonin receptors and dopamine receptors, and *in vivo* MeSH phenotypes that describe epilepsy, seizures or dementia (see middle left of Figure 14). Note that serotonin 1a and 1b receptors have one or more marketed drugs (phase 4) for the 'dementia' *in vivo* MeSH phenotype (Figure 9).
- Strong Spearman rank correlation between the hERG protein and *in vivo* MeSH phenotypes that describe epilepsy, seizures, mental disorders, bipolar disorder, depressive disorder (see right hand side of Figure 14). Note that the large number of observed Spearman rank correlations for the hERG protein (relative to other protein targets examined) are due in part because many potential drugs are tested for inhibition of hERG that would lead to cardiotoxicity.

The Spearman rank correlation approach (Figure 14) can also be used to identify additional phenotypes (side effects or toxic effects) for a specified protein target, or identify protein targets for a specified phenotype that have not yet been considered.

## 5. Difficulties

**Data selection and data gaps:** All the analyses presented in this report rely on the data selected from the ChEMBL database. The data selection was performed to provide a dataset that could be used to annotate assays to improve the database organisation, and to investigate related assays to

assess whether the method can enhance knowledge of heart and liver toxicity. The aim of this work was to demonstrate that the approach is valid using a tractable subset of the ChEMBL database. Further work is planned to apply the demonstrated approach to include other ChEMBL data and consider toxicity-related data in more depth.

**Additional data resources that could have been considered:** As part of this work, withdrawn drugs for hepato- and cardiotoxic reasons were used to identify potential toxicity-related proteins. Other sources could be used to extend this list of proteins such as drugs with a ‘black box’ warning related to cardiac or hepatic functions alteration. Other resources exist that report hepatic or cardiac side effects abstracted from books, journal or pharmacovigilance publications<sup>20</sup>. However, it is likely that a significant proportion of these drugs is already contained in the selection.

**Uneven distribution of active compounds from molecular to whole animal scale:** The selected data show a low proportion of active compounds that are tested in the binding assays, are also tested in *in vitro* functional and *in vivo* assays. One reason to explain this discrepancy is that few compounds are tested in all different types of assays from the molecular scale to the whole animal scale. For example, around 25 % of active compounds are tested in a binding, an *in vitro* functional and an *in vivo* assay in the full ChEMBL database. The uneven distribution of the number of assays between the binding, *in vitro* functional and *in vivo* assays (Table 1) will also lead to an uneven distribution of active compounds at a molecular versus whole animal scale. Additionally, note that an active compound tested in a binding assay does not necessarily show any activity in an *in vitro* functional or an *in vivo* assay. Unfortunately, this is a well-known phenomenon in drug discovery which can be due to, among other reasons, poor compound permeability or non-specific interactions in the cell.

**Active compound selection:** The definition of an active compound, versus an inactive compound, is critical. For a protein target, using the pChEMBL value is typically the most convenient way to obtain the dose-response data, giving more than 130,000 active molecules for the binding assays (Figure 4). However, many other activity types are present in the ChEMBL database and restricting to only pChEMBL values has meant ignoring a significant fraction of compounds associated with single dose data.

However, for the *in vitro* functional assays, many compounds have an activity type that does not have an associated pChEMBL value, and therefore only around 8,000 compounds were selected. To select more active compounds and, in the end, get a larger proportion of compounds shared between the binding, and *in vitro* functional assays, the constraint on the activity type could be relaxed to include single dose measurements, such as percentage of inhibition. Nevertheless, this would result in an increase of the complexity to analyse the *in vitro* functional data.

By contrast, a different approach was used to select the active compounds for the *in vivo* assays, because pChEMBL values represent a very small fraction of the *in vivo* data available. Here, all compounds were considered active if they have an associated numerical value, however, this may lead to the selection of false positives. It is noted that there are many activity types and units for the multiple *in vivo* phenotypic end-points that are described in the ChEMBL database, and it is suggested that efforts should be concentrated on the interpretability of the *in vivo* data in order to better exploit this area of the ChEMBL resource.

**Annotation:** This work represents the first time that the *in vivo* assays have been annotated in the ChEMBL database, and therefore the application of the annotation to group similar phenotypic endpoints and analyse related assay cascades marks a significant improvement of the organisation of the *in vivo* data. However, the annotation of the *in vivo* assays could be improved further as described below.

Because the reference *in vivo* assays described by Hock<sup>8</sup>, are not specific enough to precisely annotate a disease, MeSH terms have also been used as a higher-level annotation. For example, the reference assay name entitled “General Immunosuppressive Activity” was annotated with the MeSH term ‘Immune System Diseases’, but it was impossible to use a more specific MeSH term such as ‘Autoimmune Diseases’, ‘Graft vs Host Diseases’ or ‘Hypersensitivity’. Equally, reference diabetes assays such as “Alloxan Induced Diabetes”, “Streptozotocin Induced Diabetes” or “Corticosteroid Induced Diabetes” do not have a description sufficiently detailed to know which type of diabetes is being induced, so they were annotated with the MeSH terms “Diabetes Mellitus”, “Diabetes Mellitus, Type 2” and “Diabetes Mellitus, Type 1”. Therefore, further investigation, maybe using external resources, are required to allow a more accurate MeSH annotation.

Additionally, the overall annotation of the graph could be improved to distinguish the therapeutic effects of the compounds from any side effects or toxic effects if identified.

The *in vitro* functional assays were neither annotated nor pooled in this work. The attempts to group the *in vitro* functional assays according to their cell lines were not straightforward, because a given cell line can be used for different purposes. For instance, the HL60 cell line was found in different assays apparently related to different diseases (Table 2). Nevertheless, using an annotation approach similar to that carried out for the *in vivo* assays, is conceivable and might help to annotate a reasonable proportion of the *in vitro* functional assays.

**Table 2:** Example of *in vitro* functional assays performed on the cell line HL60.

Assay ChEMBL ID	Assay Description
CHEMBL871269	Induction of partial block of S phase in HL60 cells
CHEMBL939689	Inhibition of 2-[1,2-3H(N)]deoxy-D-glucose uptake in human HL60 cells at >100 uM by scintillation spectrometry in presence of sodium-free buffer
CHEMBL993750	Antioxidant activity against TPA-induced ROS production in human HL60 cells by 2',7'-dichlorodihydrofluorescein diacetate cellular-based assay

**Biological Complexity:** An important aspect that has to be taken into account is that the compounds are often active on more than one target. Promiscuity is likely to lead to side effects because the compounds binds to the therapeutic targets but also to other proteins. A selective drug will interact preferentially with its target, limiting the side effects. This can mean that assay relationships can interact and the graph becomes difficult to analyse.

Another example is where interaction of a compound with a single protein target can lead to multiple distinct phenotypes. For instance, Methotrexate, which acts as a DHFR inhibitor, is prescribed as an anticancer drug but is also used as an immune system suppressant to treat diseases like rheumatoid arthritis. Hence, the compound was found to be active for two different therapeutic areas:

- *In vitro* functional assay: “Concentration required for 50% inhibition against L1210 cells”. L1210 is a mouse lymphocytic leukemia cells used to test compounds with a tumor suppressor property;
- *In vivo* MeSH phenotype: “Adjuvant Arthritis in Rats”. This *in vivo* assay aims to induce symptoms similar to rheumatoid arthritis. Compounds of interest are then administrated to determine if they can attenuate this phenotype.

This aspect of the data makes that when looking for the assay cascades for a given compound, assays related to different indications might be mixed-up. Therefore, the finding of assay cascades related to the same indication is particularly complicated even considering the pooling and the annotation that were applied on the data.

## References

1. Gaulton, A. *et al.* The ChEMBL database in 2017. *Nucleic Acids Res.* **45**, D945–D954 (2017).
2. Atkinson, F. Deliverable Report D1.5: Package of Predictive Models. (2016).
3. Chen, M. *et al.* FDA-approved drug labeling for the study of drug-induced liver injury. *Drug Discov. Today* **16**, 697–703 (2011).
4. Papadatos, G., Gaulton, A., Hersey, A. & Overington, J. P. Activity, assay and target data curation and quality in the ChEMBL database. *J. Comput. Aided Mol. Des.* **29**, 885–896 (2015).
5. Bento, A. P. *et al.* The ChEMBL bioactivity database: an update. *Nucleic Acids Res.* **42**, D1083–D1090 (2014).
6. Nikolova, N. & Jaworska, J. Approaches to Measure Chemical Similarity— a Review. *QSAR Comb. Sci.* **22**, 1006–1026 (2003).
7. Vogel, H. G. *Drug discovery and evaluation: pharmacological assays.* (Springer, 2008).
8. Hock, F. J. *Drug Discovery and Evaluation: Pharmacological Assays.* (2016).
9. Robinson, I., Webber, J. & Eifrem, E. *Graph databases.* (O'Reilly, 2013).
10. Vorperian, V. R., Havighurst, T. C., Miller, S. & January, C. T. Adverse Effects of Low Dose Amiodarone: A Meta-Analysis. *J. Am. Coll. Cardiol.* **30**, 791–798 (1997).
11. Wolkove, N. & Baltzan, M. Amiodarone pulmonary toxicity. *Can. Respir. J. J. Can. Thorac. Soc.* **16**, 43–48 (2009).
12. Bing, R. J. & Lomnicka, M. Why do cyclo-oxygenase-2 inhibitors cause cardiovascular events? *J. Am. Coll. Cardiol.* **39**, 521–522 (2002).
13. Wang, P. & Heitman, J. The cyclophilins. *Genome Biol.* **6**, 226 (2005).
14. Elron-Gross, I., Glucksam, Y., Melikhov, D. & Margalit, R. Cyclooxygenase inhibition by diclofenac formulated in bioadhesive carriers. *Biochim. Biophys. Acta BBA - Biomembr.* **1778**, 931–936 (2008).
15. Nautiyal, K. M. & Hen, R. Serotonin receptors in depression: from A to B. *F1000Research* **6**, (2017).
16. Brisch, R. *et al.* The Role of Dopamine in Schizophrenia from a Neurobiological and Evolutionary Perspective: Old Fashioned, but Still in Vogue. *Front. Psychiatry* **5**, (2014).
17. Kuhar, M. J., Couceyro, P. R. & Lambert, P. D.  $\alpha$ - and  $\beta$ -Adrenergic Receptors. (1999).
18. Abrams, P. *et al.* Muscarinic receptors: their distribution and function in body systems, and the implications for treating overactive bladder. *Br. J. Pharmacol.* **148**, 565–578 (2006).
19. Imbrici, P., Camerino, D. C. & Tricarico, D. Major channels involved in neuropsychiatric disorders and therapeutic perspectives. *Front. Genet.* **4**, (2013).
20. Biour, M. *et al.* Drug-induced liver injury; fourteenth updated edition of the bibliographic database of liver injuries and related drugs. *Gastroenterol. Clin. Biol.* **28**, 720–759 (2004).

## ANNEX 1: Selection of protein targets

**Therapeutic targets:** We identified a set of human proteins involved in the drug mode of action. The reason behind this choice is that drugs often present side effects in combination with the therapeutic effects for which they have been approved. These proteins will be referred to as the therapeutic targets. All single proteins targeted by at least one drug have been selected from the ChEMBL database. We limited the therapeutic target set to those protein targets having drugs with at least one approved therapeutic indication. This results in a list of 153 single proteins (Annex Table 1).

**Cardiotoxic targets:** It has been demonstrated that the inhibition of some cardio-related proteins might lead to direct toxicity on this organ. For instance, the inhibition of the potassium ion channel hERG is involved in fatal cardiac arrhythmias due to repolarization disturbances of the cardiac action potential<sup>1</sup>. Many examples of target inhibition or activation that result in cardiotoxicity have been reported in the scientific literature. Recently, our group reviewed the literature to identify such protein targets and 126 of them were presented in the appendix B of the HeCaToS deliverable report D1.5 (Annex Table 1)<sup>2</sup>. In addition, a recent review reported proteins whose the link with adverse side effects has been established<sup>3</sup>. Among these proteins there were 42 proteins that interfere with cardiac function (Annex Table 1, of which 34 were found in the ChEMBL database). Another recent article highlighted some correlations between the inhibition of 30 protein kinases (Annex Table 1) and the modifications in cardiomyocyte beating rate<sup>4</sup>. Hence, by combining information from these three sources we obtained a list of 161 cardiotoxicity related protein targets.

**Hepatotoxic targets:** Unlike cardiotoxicity, there is generally no direct cause for hepatotoxicity by the interaction of molecular compounds with liver proteins. For instance, acetaminophen liver toxicity is due to its metabolite N-acetyl-p-benzoquinone imine (NAPQI) which induces a loss of glutathione with an increased formation of reactive oxygen and nitrogen species in hepatocytes undergoing necrotic changes<sup>5</sup>. Appendix B of the HeCaToS deliverable report D1.5 presents a list of potential hepatotoxic related proteins. Additionally, some other proteins have been identified using new resources from the literature to obtain a total of 94 hepatotoxicity related protein targets (Annex Table 1)<sup>3</sup>. Hence, a total of 330 unique proteins were selected as therapeutic, cardiotoxic and hepatotoxic protein targets (Annex Table 1).

### 1. Selection of withdrawn drugs and DILI-labelled drugs

This approach identified a set of 17 drugs withdrawn due to cardiotoxic reasons, and 31 drugs withdrawn due to hepatotoxic reasons. Using the approved drug labelling, 66 drugs with a description of severe drug-induced liver injury were identified and used in our study<sup>6</sup>.

### 2. Filtering of *in vitro* functional assays

From the assay description, we saw that the type of experiment performed in this assay category, is very heterogeneous. Many of them are related to the assessment of the IC<sub>50</sub> leading to toxicity of a given compound on a given cell-line. Inhibition of the growth of a cancer cell-line is also very frequent and represents a big proportion of the data collected in ChEMBL. For the purpose of this work therefore, we decided it was better to only focus on assays that could stress which molecular mechanism or pathway is studied. Knowing that a compound is toxic to a cell is valuable information, but it is of limited utility if there is no evidence on which phenomenon leads to the toxicity. This is also a way to test this approach on a relatively interpretable set of *in vitro* functional assays, before a potential application on a larger dataset.

Using a regular expression approach, we annotated each *in vitro* functional assay description to divide them in subcategories. The first category is related to cell survival. Even if the assay description is free text, the annotators tended toward using similar phrase structures for similar assays. Hence for assays coming from different articles from different years, it is frequently possible

to find the following pattern “toxicity against [cell-line name] after [time] by [test name]” or a very similar one (Table 3).

**Table 3:** Example of in vitro functional assay descriptions related to cell survival with the years they were published.

Document ChEMBL ID	Assay ChEMBL ID	Publication Year	Assay Description
CHEMBL1151266	CHEMBL998261	2005	Cytotoxicity against human A549 cells after 3 days by MTT assay
CHEMBL1151290	CHEMBL999156	1998	Cytotoxicity against human HT1080 cells after 24 hrs by MTT assay
CHEMBL1140842	CHEMBL1000149	2008	Cytotoxicity against human NCI-H838 cells after 72 hrs by MTT assay
CHEMBL1145682	CHEMBL1000731	1997	Cytotoxicity against human HT-29 cells after 7 days by MTT assay
CHEMBL1151273	CHEMBL1000883	2005	Cytotoxicity against human HepG2 cells after 48 hrs by acid phosphatase method

Similarly, many assays are related to cell division inhibition and to cancer. Hence, the terms “tumour”, “cancer”, “growth inhibition” or “antiproliferation”, in the vast majority of cases, only refer to cancer type assays (Table 4).

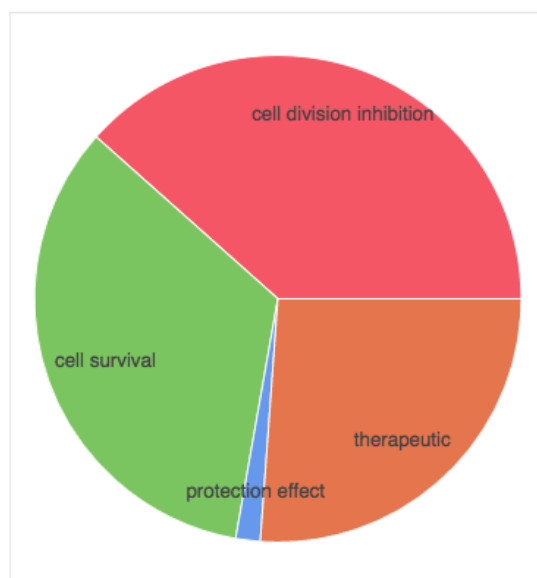
**Table 4:** Examples of in vivo assay descriptions related to inhibition of cell division with the years they were published.

Document ChEMBL ID	Assay ChEMBL ID	Publication Year	Assay Description
CHEMBL1140235	CHEMBL1000142	2008	Antiproliferative activity against human PC3 cells expressing EBP after 48 hrs by MTT assay
CHEMBL1140930	CHEMBL1000302	2009	Antiproliferative activity against human K562 cells after 48 hrs
CHEMBL1151200	CHEMBL1000745	2004	Antitumor activity against human NCI-H522 cells
CHEMBL1151219	CHEMBL996500	2004	Growth inhibition of human MCF7 cells after 48 hrs by SRB assay

Hence, using regular expression it is possible to match the assays corresponding to these categories. Because it was not possible to create categories for every assay (some types of assays are only present a couple of times), the decision was made to identify the unwanted categories and to keep only the remaining assays that will then be classified in a global ‘therapeutic’ category. Only one exception was made and it refers to assays whose aim was to evaluate the protective effect of a compound against toxicity. Indeed, the descriptions of these assays are very similar to the structure of the descriptions classified in ‘cell survival’, such as “Antihepatotoxic activity against 1.5 hrs galactosamine pretreatment-induced cytotoxicity in Wistar rat hepatocytes assessed as glutamic-pyruvic transaminase activity at 0.01 mg/ml after 30 hrs relative to control”. Taken individually, the second part of the phrase (underlined) is in agreement with the ‘cell survival’ category. Nevertheless, the mention in the first part of the phrase of “Antihepatotoxic” inversely classify this assay in the category ‘protection effect’. The repartition of the four categories identified in this work is presented in Figure 24. In the next sections, the categories ‘therapeutic’ and ‘protection effect’ will merged and will be refer as the *in vitro* functional assay set.

Looking the repartition of the categories Figure 24, one may notice that it is only possible to use about a quarter of the data (28% or 8,157 assays). The reasons why assays belong in majority to ‘cell survival’ (34%) and ‘cell division inhibition’ (38%) are several. One of them is that a majority of the journals ChEMBL curates are focused on medical chemistry. In these journals, it is frequent to find articles in which authors evaluate the toxicity of their chemical series or the effect on cancer cell-lines for them to be introduced to the scientific community. To evaluate both effects, the authors

only need to realise assays that show the effect of their compounds at a cellular scale. In our attempt to link binding assays to *in vivo* assays through *in vitro* functional assays, it is required to know what is at the origin of the cellular response.

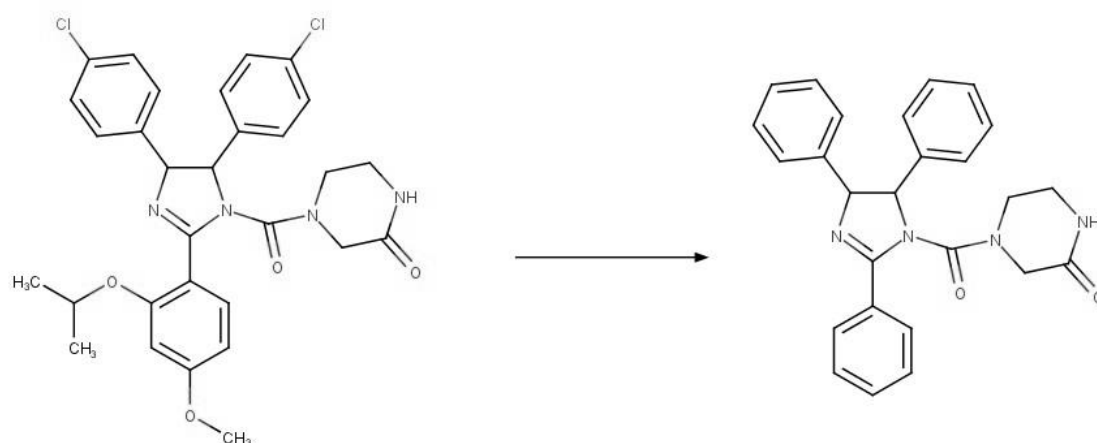


**Figure 24:** Proportions of the different categories of *in vitro* functional assays based on the assay description. The combination of the ‘therapeutic’ and ‘protection effect’ categories represent the proportion of *in vitro* functional assays used in the approach.

### 3. Assessment of the molecular similarity

**Tanimoto coefficient:** This metric is used to compare molecular fingerprints (a succession of 1 and 0, 1 indicating the presence of a specific feature and 0 its absence), by counting the number of ones that they share<sup>7</sup>. Although it is highly fingerprint dependant, molecules with a high Tanimoto coefficient ( $> 0.8$ ) usually display similar activities for the same protein targets, so this threshold was used. For this work, the well-accepted Morgan-based fingerprint with a radius of 2 and a length of 2048 bits was adopted.

**Bemis and Murcko scaffolds:** Molecular frameworks as defined by Bemis and Murcko<sup>8</sup> were used here. In their definition, they considered molecules as rings, linkers and side chains and defined molecular frameworks as the union of rings and linkers. This molecular representation is particularly helpful to group molecules that share the same molecular architecture. Side chains are often a source of diversity or specificity so, considering molecules without them allows us to consider the essence of a molecular chemotype. In our protocol, rings and linkers (set of atoms linking two rings) were retained. Exocyclic and exolinker double bonds were conserved. Atom identity was conserved on the resulting molecular framework as described on the Figure 25. This set of atoms will be referred as the molecular scaffold.



**Figure 25:** Detection of Bemis and Murcko Molecular Frameworks. The algorithm removes exocyclic and exolinker single bonds.

#### 4. Annotation of the protein targets

To keep a track of which disease area is associated with the targets, their annotation is required. This process uses the information contained in the ChEMBL database whereby each marketed drug and some preclinical candidate drugs are annotated with their indication(s). These indications use a standardised terminology called the Medical Subject Headings (MeSH), developed by the U.S. National Library of Medicine (MeSH version 2017; <https://www.nlm.nih.gov/mesh/>). MeSH provides a hierarchically-organised terminology for indexing and cataloguing of biomedical information. Hence, a drug that is described in the ChEMBL database will include information for the diseases for which it is indicated (if the information is present in the database), using the corresponding MeSH terms (e.g. indications for doxorubicin are presented in **Annex Figure 1**). The clinical phase currently associated with each indication is recorded, although this may evolve during any progression from clinical trials to marketed drug.

#### 5. Annotation of the *in vivo* assays

The key points of the approach are given below:

- Some *in vivo* reference assays described by Hock do not have any matches to the selected *in vivo* assays and therefore no text pattern has been chosen because it cannot be checked for accuracy. For example, there are no text pattern matches of the selected *in vivo* assays against the reference 'staircase test' assay for anxiolytic activity described by Hock;
- Some *in vivo* assays have more than one annotated *in vivo* reference assay described by Hock;
- Within some chapters of the *in vivo* reference assays described by Hock, it is clear that several assays can be grouped because they investigate the same phenotype. For example, multiple reference assays described by Hock consider (anti)-inflammation models (e.g. oxazolone-induced ear edema test, croton-oil ear edema test, paw edema test, pleurisy test, granuloma pouch technique, cotton wool granuloma). In such cases, an additional higher level of general annotation has been included e.g. 'general anti-inflammatory models'. This general annotation provides (i) the grouping of assays that consider a specific phenotype, and (ii) annotation of assays where the assay description may not give enough detail to assign it to an individual reference assay described in Hock, e.g. 'The compound was tested *in vivo* for locomotor activity in mice' does not give enough information to assign this assay to one specific assay, although the assay description shows that it should be grouped as one of a number of anxiolytic activity tests;
- The initial text pattern matching of selected *in vivo* assays used Vogel<sup>9</sup> as the source of the reference assays. The Vogel reference was subsequently updated, and reissued in 2016, by Hock<sup>10</sup>. In almost all cases, the description of each *in vivo* reference assay that has been used in this report is very similar between Hock and Vogel. A curation activity is planned to check (and update as required) all text pattern matches for Hock reference assays. This activity will examine

key reference assays described by Hock have not yet been curated (for functional *in vitro* and *ex vivo* assays as well as any additional *in vivo* assays);

The un annotated assays are difficult to annotate for various reasons:

- The assay description is too sparse or does not provide sufficient information to unambiguously match one reference assay described by Hock, or one group of reference assays that investigate the same phenotype. In some cases, the assay description could match two or more significantly different reference assays that use similar method to test different activity. For example, the text pattern 'lever press' may match an assay description for the 'grid shock test' that investigates central analgesic activity or it may match the 'Sidman avoidance paradigm' or the 'Geller conflict paradigm', which both investigate conditional behavioural response;
- In some cases, an assay has been inadvertently considered as *in vivo* while in fact it is a functional *in vitro* or *ex vivo* assay, and therefore no annotation can be carried out. Note that annotation of functional *in vitro* or *ex vivo* assays has not yet been carried out.

## 6. Clustering of protein targets or *in vivo* MeSH phenotypes

Construction of the protein dendrogram required that the similarity of each protein target was calculated. First, the protein classification given in ChEMBL (version 22) was extracted for each protein target. Then a simple distance matrix was calculated that was used to perform the subsequent clustering for the dendrogram.

The protein classification is given by a text description at each of eight levels. However, in some cases, one protein target was matched by more than one protein classification and therefore a sequence of rules was applied to choose the 'best' protein classification:

1. For an individual protein, if one protein classification had a more complete text description than a comparable protein classification, then the classification with more levels of text description was chosen. For example, 'acetylcholinesterase' has protein classifications of (i) level 1 = enzyme; level 2 = hydrolase, level 3 = none, etc, or (ii) level 1 = enzyme; level 2 = none, level 3 = none, etc. In this case, protein classification (i) was chosen to represent the individual protein;
2. If multiple classes of protein classification are present for one protein because there are alpha and beta subunits, then the classification distance was considered to be very close and the alpha subunit was arbitrarily chosen;
3. In a few cases, expert opinion was applied. For example, the protein classification for 'Potassium-transporting ATPase' was classed as either an enzyme, or as a transporter. In this case, the transporter classification was maintained, and the enzyme classification was dropped. Similarly, sulfonylurea receptors can be classed as either an 'ion channel' or as a 'transporter', and the 'ion channel' protein classification was chosen because it was considered to be the main biological mechanism.

Using the protein classification, each individual protein was compared against all other proteins. A distance matrix was constructed on the basis that if the protein classification for a specified level was identical for two protein targets, then the distance was set to zero, and otherwise it was set to one. The distance was summed across all classification levels to construct the overall protein distance matrix. The clustering was carried out using the SciPy hierarchical Euclidean clustering algorithm with an average distance method. Inspection of the resulting clustering was carried out to check that it was meaningful.

The process was repeated to calculate the clustering between each *in vivo* MeSH phenotype. In this case the MeSH descriptor text classification was used (e.g. 'asthma' has a MeSH descriptor of 'C08.381.495.108'. Here, the alphabetical letter was taken as the top classification level, with the following levels separated by the point character (e.g. for 'asthma', level 1 = C, level 2 = 08, level 3 = 381, level 4 = 495, level 5 = 108, level 6 = none, level 7 = none, ..., level 10 = none). In some cases,

one *in vivo* MeSH phenotype was matched by multiple MeSH classifications and the following rules were sequentially followed to choose one classification:

- If the MeSH classification level 1 descriptor of 'C' was present ('C' denotes 'Diseases') then this classification was given preference over other level 1 descriptors;
- MeSH classifications with a more complete description at each level were given preference over classifications with fewer descriptions at each level (i.e. more levels with no description);
- If multiple remaining classifications have a description for a similar number of levels, then arbitrarily the first classification was chosen.

Using the MeSH classification, each individual *in vivo* MeSH phenotype was compared against all other *in vivo* MeSH phenotypes to calculate the distance matrix. The distance was set to zero for an identical MeSH classification for two MeSH phenotypes, and otherwise it was set to one. The distance was summed across the top five classification levels to construct the overall protein distance matrix. (The use of additional MeSH classification levels did not improve the clarity of the final dendrogram since many lower levels are not described.) The same hierarchical Euclidean clustering algorithm with an average distance method was applied, and the result was inspected to check that it was reasonable.

## Annex

**Annex Table 1:** List of protein targets selected with the categories associated.

Target_name	Category			Refs
	Therapeutic	Cardiotoxic	Hepatotoxic	
3-beta-hydroxysteroid dehydrogenase/delta 5-->4-isomerase type II	X			1
3-phosphoinositide dependent protein kinase-1		X		3
AMP-activated protein kinase, alpha-1 subunit		X		3
AMP-activated protein kinase, alpha-2 subunit		X		3, 5
ATP-binding cassette sub-family A member 1	X			1
ATP-binding cassette sub-family G member 2			X	3
Acetylcholine receptor protein alpha chain		X		3
Acetylcholinesterase	X	X		1, 3, 4
Adenosine A1 receptor		X		3, 4
Adenosine A2a receptor	X	X		1, 3, 4
Adenosine A2b receptor		X		3
Adenosine A3 receptor		X		3
Adenylate cyclase type V		X		3
Aldehyde dehydrogenase	X			1
Alpha-1a adrenergic receptor	X	X		1, 3, 4
Alpha-1b adrenergic receptor		X		3, 4
Alpha-1d adrenergic receptor		X		3
Alpha-2a adrenergic receptor		X	X	3, 4
Alpha-2b adrenergic receptor		X		3, 4
Alpha-2c adrenergic receptor		X		3
Amiloride-sensitive sodium channel, ENaC	X			1
Androgen Receptor	X	X	X	1, 3
Angiotensin-converting enzyme	X	X	X	1, 3
Arachidonate 5-lipoxygenase	X			1
Aryl hydrocarbon receptor			X	3
Arylamine N-acetyltransferase 1			X	3
Arylamine N-acetyltransferase 2			X	3
Atrial natriuretic peptide receptor A		X		3
Beta-1 adrenergic receptor	X	X		1, 3, 4
Beta-2 adrenergic receptor	X	X	X	1, 3, 4
Beta-3 adrenergic receptor	X	X		1, 3
Bile acid receptor FXR	X		X	1, 3
Bile acid transporter			X	3
Bile salt export pump			X	3
Bradykinin B1 receptor		X		3
Bradykinin B2 receptor		X		3
C-C chemokine receptor type 5	X			1
C-X-C chemokine receptor type 4	X			1
CaM kinase II alpha		X		3
CaM-kinase kinase beta		X		5
Calcitonin gene-related peptide type 1 receptor		X		3

Calcium sensing receptor	X			1
Canalicular multispecific organic anion transporter 1			X	3
Canalicular multispecific organic anion transporter 3			X	3
Cannabinoid CB1 receptor	X			1
Carbamoyl-phosphate synthase [ammonia], mitochondrial	X			1
Carbonic anhydrase I	X			1
Carbonic anhydrase II	X		X	1, 3
Carbonic anhydrase IV	X			1
Carbonic anhydrase VII	X			1
Carbonic anhydrase XII	X			1
Catechol O-methyltransferase	X			1
Chloride channel protein 2	X			1
Coagulation factor X	X			1
Cyclin-dependent kinase 2		X		3
Cyclin-dependent kinase 4		X		3
Cyclin-dependent kinase 7		X		5
Cyclooxygenase-2	X	X	X	1, 3
Cysteinyl leukotriene receptor 1	X			1
Cystic fibrosis transmembrane conductance regulator	X			1
Cytochrome P450 19A1	X			1
Cytochrome P450 1A1			X	3
Cytochrome P450 1A2			X	3
Cytochrome P450 2A6			X	3
Cytochrome P450 2B6			X	3
Cytochrome P450 2C19			X	3
Cytochrome P450 2C8			X	3
Cytochrome P450 2C9			X	3
Cytochrome P450 2D6			X	3
Cytochrome P450 2E1			X	3
Cytochrome P450 3A4			X	3
Cytochrome P450 3A5			X	3
DNA topoisomerase II alpha	X			1
DOPA decarboxylase	X			1
Delta opioid receptor		X		3
Dihydrofolate reductase	X			1
Dihydroorotate dehydrogenase	X			1
Dihydropyrimidine dehydrogenase			X	3
Dipeptidyl peptidase IV	X			1
Dopamine D1 receptor	X	X		1, 3, 4
Dopamine D2 receptor	X	X		1, 3, 4
Dopamine D3 receptor	X			1
Dopamine D4 receptor		X	X	3
Dopamine transporter	X			1
Endothelin receptor ET-A		X		3, 4
Endothelin receptor ET-B		X	X	3, 4

Epidermal growth factor receptor erbB1		X		3
Equilibrative nucleoside transporter 1	X	X	X	1, 3
Equilibrative nucleoside transporter 2			X	3
Estrogen receptor alpha	X		X	1, 3
Estrogen receptor beta	X	X		1, 3
FK506-binding protein 1A	X			1
Farnesyl diphosphate synthase	X			1
Fatty acid synthase	X			1
Focal adhesion kinase 1		X		3, 5
G protein-coupled receptor kinase 5		X		3
G-protein coupled receptor kinase 2		X		3
GABA A receptor alpha-1/beta-1/gamma-2	X			1
GABA-B receptor	X			1
Gamma-amino-N-butyrate transaminase	X			1
Gastric lipase	X			1
Ghrelin receptor		X		3
Glucagon receptor		X		3
Glucocorticoid receptor	X		X	1, 3
Glutamate (NMDA) receptor subunit zeta 1		X	X	4, 3
Glutamate [NMDA] receptor subunit epsilon 1	X			1
Glutathione S-transferase Mu 1			X	3
Glutathione S-transferase Pi			X	3
Glutathione S-transferase theta 1			X	3
Glutathione reductase	X			1
Glycine receptor (alpha-1/beta)	X			1
Glycogen synthase kinase-3 alpha		X	X	3
Glycogen synthase kinase-3 beta		X		3
HERG	X	X		3
HMG-CoA reductase	X			1
Hepatocyte nuclear factor 4-alpha			X	3
Histamine H1 receptor	X	X		1, 3, 4
Histamine H2 receptor	X	X	X	1, 3, 4
Histamine H3 receptor		X		3
Ileal bile acid transporter			X	3
Inhibitor of nuclear factor kappa B kinase eps...		X		5
Inositol-1(or 4)-monophosphatase 1	X			1
Intermediate conductance calcium-activated potassium channel protein 4	X			1
Inward rectifier potassium channel 2	X	X		1, 3
Kappa opioid receptor	X	X		1, 3, 4
Kelch-like ECH-associated protein 1	X			1
LXR-alpha			X	3
LXR-beta			X	3
Leukocyte tyrosine kinase receptor		X		3
Lysosomal alpha-glucosidase	X			1
MAP kinase ERK1		X		3

MAP kinase ERK2		X	X	3
Maltase-glucoamylase	X			1
Matrix metalloproteinase 13	X			1
Matrix metalloproteinase 7	X			1
Matrix metalloproteinase 8	X			1
Matrix metalloproteinase-1	X			1
Melatonin receptor 1A		X		3
Melatonin receptor 1B		X		3
Microsomal triglyceride transfer protein	X			1
Mineralocorticoid receptor	X			1
Mitochondrial complex I (NADH dehydrogenase)	X			1
Mitochondrial glycerol-3-phosphate dehydrogenase	X			1
Mitogen-activated protein kinase kinase kinase 12		X		5
Mitogen-activated protein kinase kinase kinase 2		X		5
Mitogen-activated protein kinase kinase kinase 4		X		5
Mitogen-activated protein kinase kinase kinase 5		X		3
Mitogen-activated protein kinase kinase kinase kinase 1		X		5
Mitogen-activated protein kinase kinase kinase kinase 2		X		5
Monoamine oxidase A		X	X	3
Monoamine oxidase B	X			1
Motilin receptor			X	3
Mu opioid receptor	X	X		1, 3, 4
Multidrug and toxin extrusion protein 1			X	3
Multidrug and toxin extrusion protein 2			X	3
Multidrug resistance protein 3			X	3
Multidrug resistance-associated protein 4			X	3
Multidrug resistance-associated protein 6			X	3
Muscarinic acetylcholine receptor M1	X	X	X	1, 3, 4
Muscarinic acetylcholine receptor M2	X	X		1, 3, 4
Muscarinic acetylcholine receptor M3	X	X	X	1, 3, 4
Muscarinic acetylcholine receptor M4		X		3
Myotonin-protein kinase		X		3
NUAK family SNF1-like kinase 1		X		5
Neprilysin	X			1
Neurokinin 1 receptor	X			1
Neuronal acetylcholine receptor protein alpha-4 subunit		X		3
Neuronal acetylcholine receptor; alpha4/beta2	X			1
Neuropeptide Y receptor type 1		X		3
Niemann-Pick C1-like protein 1	X			1
Nitric-oxide synthase, endothelial		X	X	3
Norepinephrine transporter	X	X		1, 3, 4
Nuclear factor erythroid 2-related factor 2			X	3
Nuclear receptor subfamily 0 group B member 2			X	3
Nuclear receptor subfamily 1 group I member 3			X	3

Organic solute transporter subunit alpha			X	3
Orphan nuclear receptor LRH-1			X	3
P-glycoprotein 1			X	3
PI3-kinase p110-alpha subunit		X		3
PI3-kinase p110-gamma subunit		X		3
Pancreatic alpha-amylase	X			1
Pancreatic lipase	X			1
Peroxisome proliferator-activated receptor alpha	X	X	X	1, 4, 3
Peroxisome proliferator-activated receptor gamma	X		X	1, 3
Phenylalanine-4-hydroxylase	X			1
Phosphatidylinositol-5-phosphate 4-kinase type-2 beta		X		5
Phosphodiesterase 3A	X	X		1, 3, 4
Phosphodiesterase 3B		X		3
Phosphodiesterase 5A	X			1
Plasminogen	X			1
Platelet activating factor receptor		X		4
Platelet-derived growth factor receptor	X			1
Platelet-derived growth factor receptor alpha		X		3
Platelet-derived growth factor receptor beta		X		3
Potassium channel subfamily K member 10	X			1
Potassium channel subfamily K member 18	X			1
Potassium channel subfamily K member 2	X			1
Potassium channel subfamily K member 3	X	X		1, 3
Potassium channel subfamily K member 9	X			1
Potassium channel, inwardly rectifying, subfamily J, member 11	X	X		1, 3, 4
Potassium channel, inwardly rectifying, subfamily J, member 8	X			1
Potassium voltage-gated channel subfamily D member 2		X		3
Potassium-transporting ATPase	X			1
Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 2		X		3
Potassium/sodium hyperpolarization-activated cyclic nucleotide-gated channel 4	X	X		1, 3
Pregnane X receptor			X	3
Progesterone receptor	X			1
Prostanoid EP1 receptor	X			1
Prostanoid EP2 receptor	X			1
Prostanoid FP receptor	X			1
Prostanoid IP receptor	X			1
Protein kinase C alpha		X		3
Protein kinase C delta		X		5
Protein kinase C epsilon		X		5
Protein kinase C theta		X		5
Proteinase-activated receptor 1	X			1
Purinergic receptor P2Y12	X			1
Pyroglutamylated RFamide peptide receptor		X		3

Receptor protein-tyrosine kinase erbB-2		X		3
Renin	X			1
Rho-associated protein kinase 1		X		3, 5
Rho-associated protein kinase 2		X		3
Ribosomal protein S6 kinase alpha 3		X		5
Serine/threonine-protein kinase 11		X		3
Serine/threonine-protein kinase 24		X		5
Serine/threonine-protein kinase 35		X		5
Serine/threonine-protein kinase AKT		X		3
Serine/threonine-protein kinase AKT2		X		3
Serine/threonine-protein kinase Aurora-A		X		3
Serine/threonine-protein kinase Aurora-B		X		3
Serine/threonine-protein kinase Aurora-C		X		3
Serine/threonine-protein kinase B-raf		X		3
Serine/threonine-protein kinase D2		X		5
Serine/threonine-protein kinase MST1		X		5
Serine/threonine-protein kinase MST4		X		5
Serine/threonine-protein kinase PAK 1		X		5
Serine/threonine-protein kinase PAK 3		X		5
Serine/threonine-protein kinase PIM1		X		3
Serine/threonine-protein kinase PLK1		X		3
Serine/threonine-protein kinase TBK1		X		5
Serine/threonine-protein kinase ULK2		X		5
Serine/threonine-protein kinase mTOR		X		3
Serotonin 1a (5-HT1a) receptor	X			1
Serotonin 1b (5-HT1b) receptor	X	X		1, 3
Serotonin 1d (5-HT1d) receptor	X			1
Serotonin 1f (5-HT1f) receptor	X			1
Serotonin 2a (5-HT2a) receptor	X	X		1, 3
Serotonin 2b (5-HT2b) receptor	X	X		1, 3, 4
Serotonin 2c (5-HT2c) receptor	X	X		1, 3
Serotonin 3a (5-HT3a) receptor	X			1
Serotonin 4 (5-HT4) receptor	X	X		1, 3, 4
Serotonin 7 (5-HT7) receptor		X		3
Serotonin transporter	X	X		1, 3
Sigma opioid receptor	X			1
Small conductance calcium-activated potassium protein 1		X		4
Small conductance calcium-activated potassium protein 2		X		4
Small conductance calcium-activated potassium protein 3		X		4
Sodium channel protein type II alpha subunit		X		4
Sodium channel protein type IV alpha subunit	X			1
Sodium channel protein type V alpha subunit	X	X	X	1, 3
Sodium channel protein type X alpha subunit	X			1
Sodium-(potassium)-chloride cotransporter 2	X			1

Sodium/glucose cotransporter 2	X			1
Sodium/potassium-transporting ATPase alpha-1 chain		X	X	3, 4
Solute carrier family 15 member 2			X	3
Solute carrier family 22 member 1			X	3
Solute carrier family 22 member 11	X			1
Solute carrier family 22 member 12	X			1
Solute carrier family 22 member 2			X	3
Solute carrier family 22 member 6	X		X	1, 3
Solute carrier family 22 member 7			X	3
Solute carrier family 22 member 8	X			1
Solute carrier family 22 member 9			X	3
Solute carrier organic anion transporter family member 1A2			X	3
Solute carrier organic anion transporter family member 1B1			X	3
Solute carrier organic anion transporter family member 1B3			X	3
Solute carrier organic anion transporter family member 2B1			X	3
Stem cell growth factor receptor		X		3
Steroid 5-alpha-reductase 2	X			1
Succinate semialdehyde dehydrogenase	X			1
Sulfonylurea receptor 1, Kir6.2	X			1
Sulfonylurea receptor 2, Kir6.2	X			1
Sulfotransferase 1A1			X	3
Synaptic vesicle glycoprotein 2A	X			1
Synaptic vesicular amine transporter	X			1
Thiazide-sensitive sodium-chloride cotransporter	X			1
Thiopurine S-methyltransferase			X	3
Thrombin	X	X		1, 3
Thrombopoietin receptor	X			1
Thromboxane A2 receptor		X		3
Thymidylate synthase	X			1
Thyroid hormone receptor	X			1
Thyroid peroxidase	X			1
Tissue-type plasminogen activator	X			1
Toll-like receptor 7	X			1
Toll-like receptor 9	X			1
Type I iodothyronine deiodinase (Type-I 5'-deiodinase) (DIOI) (Type 1 DI) (5DI)	X			1
Type-1 angiotensin II receptor	X	X		1, 3
Tyrosine 3-hydroxylase	X	X		1, 3
Tyrosine-protein kinase ABL		X		3
Tyrosine-protein kinase FER		X		5
Tyrosine-protein kinase JAK2		X		3
Tyrosine-protein kinase ZAP-70		X		5
UDP-glucuronosyltransferase 1-1			X	3
UDP-glucuronosyltransferase 2B15			X	3

UDP-glucuronosyltransferase 2B17			X	3
UDP-glucuronosyltransferase 2B7			X	3
Urotensin II receptor		X		3
Vanilloid receptor	X			1
Vascular endothelial growth factor receptor 1		X		3
Vascular endothelial growth factor receptor 2		X		3
Vascular endothelial growth factor receptor 3		X		3
Vasopressin V1a receptor	X	X		1, 3, 4
Vasopressin V1b receptor		X		3
Vasopressin V2 receptor	X			1
Vitamin D receptor	X		X	1, 3
Vitamin k epoxide reductase complex subunit 1 isoform 1	X			1
Voltage-gated L-type calcium channel alpha-1C subunit		X		3, 4
Voltage-gated N-type calcium channel alpha-1B subunit	X			1
Voltage-gated T-type calcium channel alpha-1G subunit		X		3
Voltage-gated T-type calcium channel alpha-1H subunit		X		3
Voltage-gated potassium channel beta subunit Mink		X		3
Voltage-gated potassium channel subunit Kv1.4		X		3
Voltage-gated potassium channel subunit Kv1.5		X		3
Voltage-gated potassium channel subunit Kv3.1		X		3
Voltage-gated potassium channel subunit Kv4.3		X		3
Voltage-gated potassium channel subunit Kv7.1		X		3
Xanthine dehydrogenase	X			1
cAMP-dependent protein kinase alpha-catalytic subunit		X		5
cGMP-dependent protein kinase 1 beta		X		3, 5

**Annex Table 2:** List of drugs selected for toxicity reasons.

<b>mol_chemblid</b>	<b>pref_name</b>	<b>category</b>
CHEMBL106	FLUCONAZOLE	DILI labeling
CHEMBL107	COLCHICINE	DILI labeling
CHEMBL108	CARBAMAZEPINE	DILI labeling
CHEMBL109	VALPROIC ACID	DILI labeling
CHEMBL1094	FELBAMATE	DILI labeling
CHEMBL115	INDINAVIR	DILI labeling
CHEMBL1185568	DITHIAZANINE IODIDE	withdrawn for cardiotoxicity reasons
CHEMBL1201288	DANTROLENE	DILI labeling
CHEMBL121	ROSIGLITAZONE MALEATE	withdrawn for cardiotoxicity reasons
CHEMBL122	ROFECOXIB	withdrawn for cardiotoxicity reasons
CHEMBL12713	SERTINDOLE	withdrawn for cardiotoxicity reasons
CHEMBL129	ZIDOVUDINE	DILI labeling
CHEMBL1297	FENOPROFEN	DILI labeling
CHEMBL1324	TOLCAPONE	DILI labeling, withdrawn for liver toxicity reasons
CHEMBL1341	METHOXYFLURANE	withdrawn for liver toxicity reasons
CHEMBL1371	CHLORZOXAZONE	DILI labeling
CHEMBL1380	ABACAVIR	DILI labeling
CHEMBL139	DICLOFENAC	DILI labeling
CHEMBL141	LAMIVUDINE	DILI labeling
CHEMBL141305	CYCLOFENIL	withdrawn for liver toxicity reasons
CHEMBL1419	SIBUTRAMINE HYDROCHLORIDE	withdrawn for cardiotoxicity reasons
CHEMBL1425	MERCAPTOPURINE	DILI labeling
CHEMBL1479	DANAZOL	DILI labeling
CHEMBL1481	GLIMEPIRIDE	DILI labeling
CHEMBL1515	METHIMAZOLE	DILI labeling
CHEMBL1518	PROPYLTHIOURACIL	DILI labeling
CHEMBL1560	CAPTOPRIL	DILI labeling
CHEMBL15770	SULINDAC	DILI labeling
CHEMBL1643	RIBAVIRIN	DILI labeling
CHEMBL17157	TERFENADINE	withdrawn for cardiotoxicity reasons
CHEMBL175247	ORLISTAT	DILI labeling
CHEMBL182	GANCICLOVIR	DILI labeling
CHEMBL222559	TIPRANAVIR	DILI labeling
CHEMBL254857	FIPEXIDE	withdrawn for liver toxicity reasons
CHEMBL267744	TICRYNAFEN	withdrawn for liver toxicity reasons
CHEMBL27193	DILEVALOL	withdrawn for liver toxicity reasons
CHEMBL273575	NOMIFENSINE	withdrawn for liver toxicity reasons
CHEMBL282724	SITAXENTAN SODIUM	withdrawn for liver toxicity reasons
CHEMBL295698	KETOCONAZOLE	DILI labeling
CHEMBL340978	BENOXAPROFEN	withdrawn for liver toxicity reasons
CHEMBL341812	IBUFENAC	withdrawn for liver toxicity reasons
CHEMBL363295	TERODILINE	withdrawn for cardiotoxicity reasons
CHEMBL374478	RIFAMPICIN	DILI labeling
CHEMBL400599	BENFLUOREX	withdrawn for cardiotoxicity reasons
CHEMBL404108	LUMIRACOXIB	withdrawn for liver toxicity reasons
CHEMBL408	TROGLITAZONE	withdrawn for liver toxicity reasons
CHEMBL409	BICALUTAMIDE	DILI labeling
CHEMBL429	LABETALOL	DILI labeling
CHEMBL443	SULFAMETHOXAZOLE	DILI labeling
CHEMBL445	NORTRIPTYLINE	DILI labeling

CHEMBL459	METHYLDOPA	DILI labeling
CHEMBL467	HYDROXYUREA	DILI labeling
CHEMBL476	DACARBAZINE	DILI labeling
CHEMBL477772	PAZOPANIB	DILI labeling
CHEMBL479	THIORIDAZINE HYDROCHLORIDE	withdrawn for cardiotoxicity reasons
CHEMBL483	TENOFOVIR	DILI labeling
CHEMBL490	PAROXETINE	DILI labeling
CHEMBL531	PERGOLIDE MESYLATE	withdrawn for cardiotoxicity reasons
CHEMBL535	SUNITINIB	DILI labeling
CHEMBL550348	DEFERASIROX	DILI labeling
CHEMBL554	LAPATINIB	DILI labeling
CHEMBL57	NEVIRAPINE	DILI labeling
CHEMBL572	NITROFURANTOIN	DILI labeling
CHEMBL573	NIACIN	DILI labeling
CHEMBL583	GREPAFLOXACIN HYDROCHLORIDE	withdrawn for cardiotoxicity reasons
CHEMBL588119	SULOCTIDIL	withdrawn for liver toxicity reasons
CHEMBL603	ZAFIRLUKAST	DILI labeling
CHEMBL622	ETODOLAC	DILI labeling
CHEMBL623	NEFAZODONE	DILI labeling
CHEMBL633	AMIODARONE	DILI labeling
CHEMBL637	VENLAFAXINE	DILI labeling
CHEMBL64	ISONIAZID	DILI labeling
CHEMBL76370	TEGASEROD MALEATE	withdrawn for cardiotoxicity reasons
CHEMBL8	CIPROFLOXACIN	DILI labeling
CHEMBL806	FLUTAMIDE	DILI labeling
CHEMBL822	TERBINAFINE	DILI labeling
CHEMBL83	TAMOXIFEN	DILI labeling
CHEMBL853	ZALCITABINE	DILI labeling
CHEMBL865	VALDECOXIB	withdrawn for cardiotoxicity reasons
CHEMBL92	DOCETAXEL HYDRATE	DILI labeling
CHEMBL95	TACRINE	DILI labeling
CHEMBL957	BOSENTAN	DILI labeling
CHEMBL960	LEFLUNOMIDE	DILI labeling
CHEMBL964	DISULFIRAM	DILI labeling
CHEMBL991	STAVUDINE	DILI labeling

MESH Heading	MESH ID	EFO ID	EFO Term	Max phase for indication	References
NEOPLASMS	<a href="#">D009369</a>	<a href="#">EFO:0000311</a>	CANCER	4	<a href="#">ClinicalTrials</a>
NEOPLASMS	<a href="#">D009369</a>	<a href="#">EFO:0000616</a>	NEOPLASM	4	<a href="#">ATC</a> <a href="#">ClinicalTrials</a> <a href="#">ClinicalTrials</a>
BREAST NEOPLASMS	<a href="#">D001943</a>	<a href="#">EFO:0000305</a>	BREAST CARCINOMA	4	<a href="#">ClinicalTrials</a> <a href="#">ClinicalTrials</a>
PROSTATIC NEOPLASMS	<a href="#">D011471</a>	<a href="#">EFO:0001663</a>	PROSTATE CARCINOMA	3	<a href="#">ClinicalTrials</a> <a href="#">ClinicalTrials</a>
SARCOMA, KAPOSI	<a href="#">D012514</a>	<a href="#">EFO:0000558</a>	KAPOSI'S SARCOMA	3	<a href="#">ClinicalTrials</a>
URINARY BLADDER NEOPLASMS	<a href="#">D001749</a>	<a href="#">EFO:0000292</a>	BLADDER CARCINOMA	3	<a href="#">ClinicalTrials</a>
SMALL CELL LUNG CARCINOMA	<a href="#">D055752</a>	<a href="#">EFO:0000702</a>	SMALL CELL LUNG CARCINOMA	3	<a href="#">ClinicalTrials</a>
ENDOMETRIAL NEOPLASMS	<a href="#">D016889</a>	<a href="#">EFO:0004230</a>	ENDOMETRIAL NEOPLASM	3	<a href="#">ClinicalTrials</a>
OVARIAN NEOPLASMS	<a href="#">D010051</a>	<a href="#">EFO:0003893</a>	OVARIAN NEOPLASM	3	<a href="#">ClinicalTrials</a> <a href="#">ClinicalTrials</a>
SARCOMA	<a href="#">D012509</a>	<a href="#">EFO:0000691</a>	SARCOMA	3	<a href="#">ClinicalTrials</a> <a href="#">ClinicalTrials</a>

**Annex Figure 1:** Indications for doxorubicin, as displayed on ChEMBL website compound card.

**Annex Table 3:** List of HeCaToS compounds with their ChEMBL identifier. The toxicity category identified by the HeCaToS project has been shown for each compound.

PARENT PREF NAME	PARENT ChEMBL ID	HECATOS CATEGORY
ACETAMINOPHEN	ChEMBL112	HEPATOTOXIC
AMIODARONE	ChEMBL633	CARDIOTOXIC
AZATHIOPRINE	ChEMBL1542	HEPATOTOXIC
CELECOXIB	ChEMBL118	CARDIOTOXIC
CLAVULANIC ACID	ChEMBL777	HEPATOTOXIC
CYCLOPHOSPHAMIDE	ChEMBL88	CARDIOTOXIC
CYCLOSPORINE	ChEMBL160	HEPATOTOXIC
DAUNORUBICIN	ChEMBL178	CARDIOTOXIC
DICLOFENAC	ChEMBL139	HEPATOTOXIC
DOCETAXEL	ChEMBL92	CARDIOTOXIC
DOXORUBICIN	ChEMBL53463	CARDIOTOXIC
EPIRUBICIN	ChEMBL417	CARDIOTOXIC
ERYTHROMYCIN	ChEMBL532	HEPATOTOXIC
FLUOROURACIL	ChEMBL185	CARDIO/HEPATOTOXIC
IDARUBICIN	ChEMBL1117	CARDIOTOXIC
INFLIXIMAB	ChEMBL1201581	CARDIOTOXIC
IRINOTECAN	ChEMBL481	HEPATOTOXIC
ISONIAZID	ChEMBL64	HEPATOTOXIC
LAPATINIB	ChEMBL554	CARDIOTOXIC
METHOTREXATE	ChEMBL34259	HEPATOTOXIC
MITOXANTRONE	ChEMBL58	CARDIOTOXIC
PACLITAXEL	ChEMBL428647	CARDIOTOXIC
PHENOBARBITAL	ChEMBL40	HEPATOTOXIC
PHENYTOIN	ChEMBL16	HEPATOTOXIC
PIROXICAM	ChEMBL527	HEPATOTOXIC
PRAVASTATIN	ChEMBL1144	HEPATOTOXIC
RIFAMPICIN	ChEMBL374478	HEPATOTOXIC
SIMVASTATIN	ChEMBL1064	HEPATOTOXIC
SORAFENIB	ChEMBL1336	CARDIOTOXIC
TETRACYCLINE	ChEMBL1440	HEPATOTOXIC
VALPROIC ACID	ChEMBL109	HEPATOTOXIC

## References

1. Raschi, E., Ceccarini, L., De Ponti, F. & Recanatini, M. hERG-related drug toxicity and models for predicting hERG liability and QT prolongation. *Expert Opin. Drug Metab. Toxicol.* **5**, 1005–1021 (2009).
2. Atkinson, F. Deliverable Report D1.5: Package of Predictive Models. (2016).
3. Lynch, J. J., Van Vleet, T. R., Mittelstadt, S. W. & Blomme, E. A. G. Potential functional and pathological side effects related to off-target pharmacological activity. *J. Pharmacol. Toxicol. Methods* (2017). doi:10.1016/j.vascn.2017.02.020
4. Lamore, S. D. *et al.* Deconvoluting Kinase Inhibitor Induced Cardiotoxicity. *Toxicol. Sci.* (2017). doi:10.1093/toxsci/kfx082
5. Hinson, J. A., Roberts, D. W. & James, L. P. Mechanisms of Acetaminophen-Induced Liver Necrosis. in *Adverse Drug Reactions* (ed. Uetrecht, J.) **196**, 369–405 (Springer Berlin Heidelberg, 2010).
6. Chen, M. *et al.* FDA-approved drug labeling for the study of drug-induced liver injury. *Drug Discov. Today* **16**, 697–703 (2011).
7. Bajusz, D., Rácz, A. & Héberger, K. Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *J. Cheminformatics* **7**, (2015).
8. Bemis, G. W. & Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **39**, 2887–2893 (1996).
9. Vogel, H. G. *Drug discovery and evaluation: pharmacological assays*. (Springer, 2008).
10. Hock, F. J. *Drug Discovery and Evaluation: Pharmacological Assays*. (2016).