



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 D1.5:
Package of Predictive Models**

Edited by

Francis Atkinson, EMBL-EBI

Work package 1

Due date of deliverable: M21

Actual submission date: M24

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 Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Contributions to deliverable - Internal review procedure

Deliverable produced by:	Date:
Francis Atkinson, EMBL-EBI	August, 2015
Anne Hersey, EMBL-EBI	August, 2015
Deliverable internally reviewed by:	Date:
Jos Kleinjans	Sept 8, 2015

Contents

Publishable Summary.....	3
Objectives	3
Introduction	4
Results.....	6
Consolidation of list of targets taken from the literature.....	6
Mapping targets to ChEMBL database	8
Species	10
Active Compounds	11
Compound Sizes.....	14
Target Classes	15
Compound diversity.....	18
Modelling: strategy	22
Modelling: results	23
Difficulties	28
References	29
D1.5 Appendix A.....	30
References	50

PUBLISHABLE SUMMARY

This report describes the development, using [ChEMBL](#) data, of a package of predictive models for toxicology-associated targets described in a previous EMBL-EBI deliverable (D9.2). These targets are biomolecules that can cause a toxic response when they interact with a xenobiotic. The models are intended to be useful for two main reasons. Firstly, they aim to identify molecules that might bear toxicological risks due to direct interaction with the targets of interest. Second, they are intended to be incorporable into 'higher-level' models, where interactions are with a biochemical pathway or involve the broader ADME properties of the molecule. The workflow for developing models is presented, and various aspects of their performance are discussed.

OBJECTIVES

One of the objectives of HeCaToS Work Package 1 (WP1) is "Develop library of models for all suitable toxicity targets".

This report describes work performed at EMBL-EBI to address this objective. The HeCaToS *tasks* associated with the objective involve various aspects of modelling activities of toxicological interest:

- T1.6 - Develop models of metabolite/target/pathway/ADME activities using ChEMBL data to be used for integration with PBPK modelling under WP4 (OPTI/EMBL);
- T1.8 - Develop naïve-Bayes ECFPP and similarity ensemble approach models for target prediction for a molecule (EMBL);
- T1.9 - Develop and validate library of models for all suitable toxicity targets (EMBL).

The activities discussed here are seen as part of an on-going effort to accomplish these tasks and thus fulfil the objective. Thus, the focus has not just been on producing models, but rather on establishing robust workflows for pre-processing data and model generation. These have been designed to be straightforward to run, such that models can be easily regenerated as new data becomes available and as further targets of interest are identified. A further goal is that the workflows facilitate exploration of alternative modelling strategies. Together these principles are intended to ensure that the quality and scope of the models can be improved as the project proceeds.

Further work will be described in future WP1 reports, most specifically HeCaToS WP1 Deliverable Report D1.7: 'Report on predictions for library of toxicity related targets' (M48).

INTRODUCTION

This report describes the development of a package of predictive models based on the targets described in the EMBL-EBI Deliverable D9.2: 'Report on reference set of toxicology associated targets'. An updated version of that document is included here as Appendix A, and the reader is assumed to be familiar with the contents. An Excel Workbook containing the list of targets discussed in D9.2 is included here as Appendix B.

Briefly, a 'target' is used in this document to mean a biomolecule, generally a protein that can mediate a toxic response when it interacts with a xenobiotic, which could be a drug or environmental chemical or a metabolite thereof. These are 'low-level' interactions, possibly comprising the molecular initiating event (MIE) [1] of an adverse outcome pathway (AOP) [2]. Other, 'higher-level' definitions of target are also possible and correspond to other stages of the AOP; examples include biochemical pathways, organelles, cells, organs or even whole organisms. The prediction of the effects of xenobiotics on these more complex entities is also of interest to HeCaToS, and will be the subject of other deliverables.

It should be noted that the xenobiotics of interest here are small organic molecules. Other bioactive entities, such as protein therapeutics (*e.g.* antibodies) or inorganic/organometallic compounds (*e.g.* platin drugs) are not considered. The cheminformatics tools available do not currently handle such entities well, and the conceptual structure-activity relationships that implicitly underlie the modelling strategy used do not really apply.

The ChEMBL database (<https://www.ebi.ac.uk/chembl>) [3] is the source of all the target/small-molecule activity data used in this analysis. It is freely distributable, well curated and contains data on a wide variety of targets of interest for toxicological modelling. However, there are some targets of interest for which the amount of data is small or even non-existent. This seems to be a particular problem for data on transporters, ion channels (other than hERG) and Phase II xenobiotic metabolising enzymes. One possible reason for this is that interest in a target class is relatively recent and data generated within pharmaceutical companies has not had time to reach the medicinal chemistry literature. Another reason might be that the assays are low-throughput and therefore most often performed at a later stage in the drug discovery pipeline. This can mean that they are reported in journals other than those routinely covered by ChEMBL, which has historically focussed on the earlier stages of the pipeline. As data for some of these targets could be of particularly high value to HeCaToS, effort has been expended on investigating possible strategies for enriching the database in a focussed manner.

The modelling strategy adopted here is purely qualitative; molecules are to be classified as either active or inactive, and there is no attempt at present to provide a quantitative estimate of activity. This is a purely pragmatic decision, partly designed to avoid (for now) the need for time-consuming data curation, and also in view of the fact that many targets of interest have relatively few data points derived from a heterogeneous set of assays. In some cases, even a clean separation into active and inactives training sets can be problematic, and any quantitative prediction is unfeasible.

A further issue that is not addressed in detail here is that of differing assay endpoints. For example, GPCR or nuclear receptor ligands can be full agonists, antagonists or somewhere on the spectrum between these extremes. So-called functional assays measure these specific activities, whereas others simply measure binding to the receptor. In the present analysis, no distinction is made between these assay types, and 'activity' is thus any engagement of the receptor. The justification for this is, again, pragmatic: the nature of the data is such that, to allow an unambiguous assignment of the activity type, a degree of manual curation would be required that is not plausible at this time. Such an assignment might not even be possible with, for example, partial agonists; furthermore, due to overlapping binding sites, rather subtle modifications to a compound can change the balance

between agonism and antagonism significantly. Thus, at least for the current qualitative analysis, ignoring the endpoint seems supportable for now.

A related, though distinct, issue occurs with xenobiotic metabolising enzymes such as the CYP450s. In some cases, it may not be inhibition of the enzyme that is of interest, but rather its induction. In this case, although the readout might involve changes in enzymatic activity, it is the compound binding to the nuclear receptor that controls expression of the enzyme that is actually being measured. SAR for the enzyme itself and the coupled NR need not be related, and confusion can result where annotation of the data is inadequate. However, such cases are rare and this is not perceived to be a problem at present.

Although the focus here is on qualitative models, the quantitative predictions of activity for that subset of targets for which the data will support them would be valuable. However, the decision was made that EMBL-EBI would prioritise breadth of coverage, hence the use of qualitative (*i.e.* classification) over more sophisticated quantitative (*e.g.* regression) models.

The modelling method used here is Multi-Category Naïve Bayes (MCNB) [4] using ECFP4 fingerprints [5]. This combination was chosen due the ease of implementing panel predictions using this technique, and because it had been used successfully in a previous ChEMBL-based target-prediction exercise [6]. Although this classifier is simple to run and has produced useful results in the past it does have limitations. For example, only active compounds are used in the training procedure, all compounds not being active against a particular target being assumed to be inactive. Thus, for example, information about inactives closely related to the actives is lost. An obvious next step will be to implement individual classification models for each target and see whether any performance improvement is obtained. Similarly, other fingerprinting techniques need to be examined, such as those incorporating pharmacophoric information.

As stated above, the focus has not just been on producing the package of models, but rather on establishing robust workflows for the required data pre-processing and generation of models. These have been designed to be simple to run, with as little manual intervention as possible required. Thus, when a new version of ChEMBL is released, the models will be regenerated, automatically incorporating such new data as is available. Similarly, if new biomolecular targets of toxicological interest are identified, they may be added to the panel in a straightforward manner. This has led to a preference for simplicity in the modelling process, at least for this first iteration, and to a pragmatic approach in the preparation of data for modelling. It is possible that more sophisticated approaches might lead to somewhat improved results. For example, Random Forest classifiers or Support Vector Machines might outperform the Naïve Bayes classifiers currently employed and the workflow is also intended to facilitate experimentation with the modelling techniques used. Similarly, a more 'hands-on' approach to data curation and pre-processing might increase the number of useable data points. However, such activities are particularly time-consuming, and thus must be balanced against other areas of interest.

The models described herein are intended to be used in two main ways. Firstly, they aim to identify molecules that might have toxicological risks associated with them due to their direct interaction with the targets of interest. Secondly, they are intended to be incorporated into 'higher-level' models, where the interactions of interest are primarily with, for example, biochemical pathways. If a xenobiotic is known (or predicted) to interact with a biomolecule that is a member of a pathway, it is plausible that the xenobiotic may perturb that pathway and hence such cellular processes as are dependent upon it. As well as 'low-level' assays measuring the interaction of small molecules with biomolecules, the ChEMBL database also contains data from functional and phenotypic assays. The hope is that this 'high-level' data will be used to both build models of these endpoints directly and also to validate hierarchical models based on the kind of biomolecular-level model described herein. Analysis of pathways might also suggest new molecular targets for inclusion in the panel. Reactome

[7] is an excellent source of pathway data and is being used as the source of pathway information for this project.

The choice of software used in this work was determined by two factors: that the tools are suitable for the task in hand and that they are free for all to use. The latter is important as it ensures others could replicate the work if necessary. For example, Pipeline Pilot is a powerful platform for data processing and modelling; however, it is proprietary software and was thus not considered suitable. By contrast, the Python data-science ecosystem [8] includes powerful tools for data manipulation (pandas) and modelling (scikit-learn), all of which are open-source. In addition the IPython Notebook, which allows the interweaving of code, HTML and images such as plots provides an excellent way of both performing and documenting workflows. There is also excellent integration with the RDKit [9], which is used for all cheminformatics requirements.

The workflow may be broadly broken down conceptually into the following stages, which are discussed in the results section below:

- Consolidation of lists of targets taken from the literature;
- Mapping these targets to ChEMBL database;
- Retrieval and filtering of activity data from ChEMBL;
- Building of panel of models;
- Testing and analysis of models.

IPython Notebooks provide both the documentation and reference implementation of the models and are available for download or inspection at https://github.com/flatkinson/tox_models. This repository will be kept updated as improvements are made.

A REST-based web service that will allow programmatic access to the panel of models is currently under development and will be available soon.

Please note that, for practical reasons, the IPython Notebooks do not yet map exactly to the scheme described above. They are being refactored so they conform to it more closely. Note also that the figures in this document are taken from the Notebooks; thus, if any figure is unclear, the original is available for inspection.

RESULTS

Consolidation of list of targets taken from the literature

The first step was to consolidate the various lists of targets related to hepato- and cardiotoxicity compiled in the HeCaToS Deliverable Report D9.2 (*N.B.* see Appendix A for an updated version of that document). The lists included there are from a variety of sources and naturally overlap to a certain extent. In addition, targets are recorded in a variety of ways; often a HUGO gene symbol [10] is used, but sometimes a protein name or other identifier is used. These are frequently non-standard, and occasionally some research was required in order to unambiguously identify the intended target.

To progress with de-duplication and linking to ChEMBL activity, it was necessary to assign a common identifier to each target. The HUGO gene symbol was chosen for this task, although other choices would have been possible (such as UniProt Accession Numbers, Entrez Gene IDs *etc.*). The gene symbol was chosen as it was already present in many cases, and provides a widely used and convenient identifier (for example, conversion to other identifiers is generally straightforward). All targets were thus assigned gene symbols; Appendix B contains versions of all the tables in Appendix A but with a 'Gene' column containing the symbol added when necessary.

Note that in some cases the original literature contained subtle misassignments or typographical errors. An example of the former is Table 5, where 'PDK1' is given as the gene symbol for Phosphoinositide-Dependent Kinase-1I whereas the appropriate symbol is 'PDK1'. PDK1 is actually the symbol for Pyruvate Dehydrogenase Kinase, isozyme 1; although they are both kinases, it is clear from the article that PDK1 is correct.

Note that while errors have been fixed in both appendices, if there is a discrepancy then the Workbook (Appendix B) is to be regarded as the definitive source.

There are several issues to be dealt with in assigning gene symbols to targets. One example is where a target corresponds to multiple isoforms; for example, in Table 1 one target is given as 'Acetylcholine receptor subunit $\alpha 1$ or $\alpha 4$ (CHRNA1 or CHRNA4)'. In the context of the source document, this means that one or the other should be included in a screening panel. In the present context, both must be included as it is not clear at this stage which would be more appropriate (*i.e.* be somehow 'more representative' or perhaps simply have more data associated with it).

The inclusion of multiple isoforms, either given explicitly or included because the source does not specify which isoforms are meant, occurs in a number of cases. For example, Table 5 (kinases) includes several examples of both: targets include 'AKT1, 2 or 3' and 'GSK3 α/β ' but also 'CDKs' and 'Aurora kinases'. The Aurora kinase family contains three members and all were included. The CDKs are a large family, however, and after inspection of the source document only CDK2 and CDK4 were included. The inclusion of multiple isoforms is not considered to be a problem as such, although it might arguably be said to introduce a certain amount of redundancy into the panel.

ChEMBL also contains data for which a set of isoforms (*e.g.* ChEMBL2111445, 'Muscarinic acetylcholine receptor M2 and M4') or an entire family (*e.g.* ChEMBL2094109, 'Muscarinic acetylcholine receptor') is named as the target. This can be because it is not known (or is not clear in the source) which isoforms induce a response or because it is known that multiple isoforms are involved. Because of the difficulties in interpreting this data it is not used at present, although its use will be investigated in future.

Another issue is where the target is a protein complex, and therefore maps to multiple gene symbols. Again in Table 1 is the target 'Potassium voltage-gated channel KQT-like member 1 (KCNQ1) and minimal potassium channel MinK (KCNE1)'. Here, KCNQ1 encodes potassium channel KvLQT1, while KCNE1 encodes an ancillary protein MinK that modifies the activity of the channel. Together, they give rise to the IKs current in cardiac myocytes. Thus, strictly speaking, it is the complex they form that is the target of interest. However, in other sources (*e.g.* that from which Table 4 is taken), the IKs current is identified with KCNQ1, which is reasonable given that it is clearly the central component of the complex. Further, a preliminary inspection of the ChEMBL data shows that, while there is data for both KvLQT1 and the KvLQT1/MinK complex, the latter is negligible (four data points, all inactive). Thus, in this case, nothing is lost by using the main component alone to represent the complex as a whole. It should be noted that this is done also in the previous example, where the α -subunit of the acetylcholine receptor (*i.e.* CHRNA1 or CHRNA4) is used to stand in for the complete receptor complex.

One potentially very complex example is ATPase (Na^+/K^+), the sodium-potassium pump. This is composed of α - and β -subunits, each of which has multiple possible isoforms that can combine in various ways. However, small-molecule activity data vs. this complex appears to be rare and, in fact, the target does not currently appear in ChEMBL.

A preliminary analysis of the data showed that all complexes of interest here can be represented by a main subunit (or isoforms thereof as with the nAChRs) without ambiguity or significant loss of

data. This strategy has therefore been adopted in practice as it simplifies matters considerably, although it is certainly conceivable that in future it might be necessary to adopt a more realistic (and hence complicated) naming scheme for the targets.

Hindsight suggests that it is not really surprising this simplification is possible, as data against more complex targets is often more difficult to generate in practice and is therefore less common. Furthermore, *in vitro* assays against such targets often use simplified versions of the complex and/or use the main component as shorthand for the complete entity when reporting data in the literature.

There is, of course, a price to be paid for this pragmatic strategy, in that some of the complexity of the underlying biology is hidden. In some cases, such as with KvLQT1/MinK, there is so little data for the complex relative to its main component that the issue is moot. In others cases, there is non-trivial amounts of data against both the main component and the complex or complexes including that component. In such cases the best course of action is not always obvious: this is discussed more in the following section.

One issue that should be mentioned is that there is an implicit assumption that small molecule xenobiotics may interact with the target to modulate its activity. In some cases, this modulation may not be direct, and this might not be immediately obvious from the assay description available in ChEMBL. One example is Nrf2 ([NFE2L2](#)), a transcription factor that has a complex mechanism of regulation (highly unlike the ligand-modulated nuclear receptors, for example. As the amount of data for this target is small and it is thus not a candidate for modelling (see below), this is not a problem at the moment. However, it highlights the need for careful curation of targets, and perhaps exclusion of those where small-molecule interactions are unlikely to be important.

Mapping targets to ChEMBL database

Once HUGO gene symbols were assigned to all targets, the list was collated and duplicates eliminated; this gave a total of 215 targets. This list of unique gene symbols was then mapped to ChEMBL targets.

ChEMBL targets may be single proteins, protein complexes or protein families. The latter is used in cases where the precise isoform is not specified in the literature: an example is 'Glycogen synthase kinase-3' ([CHEMBL2095188](#)), which is assigned if the GSK3 isoform, 'Glycogen synthase kinase-3 alpha' ([CHEMBL1075224](#)) or 'Glycogen synthase kinase-3 beta' ([CHEMBL262](#)), is not known. However, it was decided that, given the decision to use gene symbols as the core identifier, protein families would not be included in this exercise. Thus only 'single protein' and 'protein complex targets' were considered.

The mapping was performed using the 'component_synonyms' table in ChEMBL, which contains the HUGO gene symbols corresponding to each protein chain. Single-protein targets have a single component, while complex targets have several. For each symbol, all targets having that protein as a component were retrieved.

Of the initial 215, 20 did not appear in ChEMBL. These symbols correspond to ten distinct targets, including the sodium-potassium pump and three cardiac ion channels. These four targets correspond to 14 of the lost symbols; the rest are almost all transporters.

In most cases, only a single ChEMBL target exists per symbol. In a significant minority of cases (18 of the remaining 195), however, multiple ChEMBL targets exist. For example, for CDK2 there are various entities in ChEMBL with non-trivial amounts of data associated with them; these are shown in Table 1, where N is the number of distinct active compounds associated with each target.

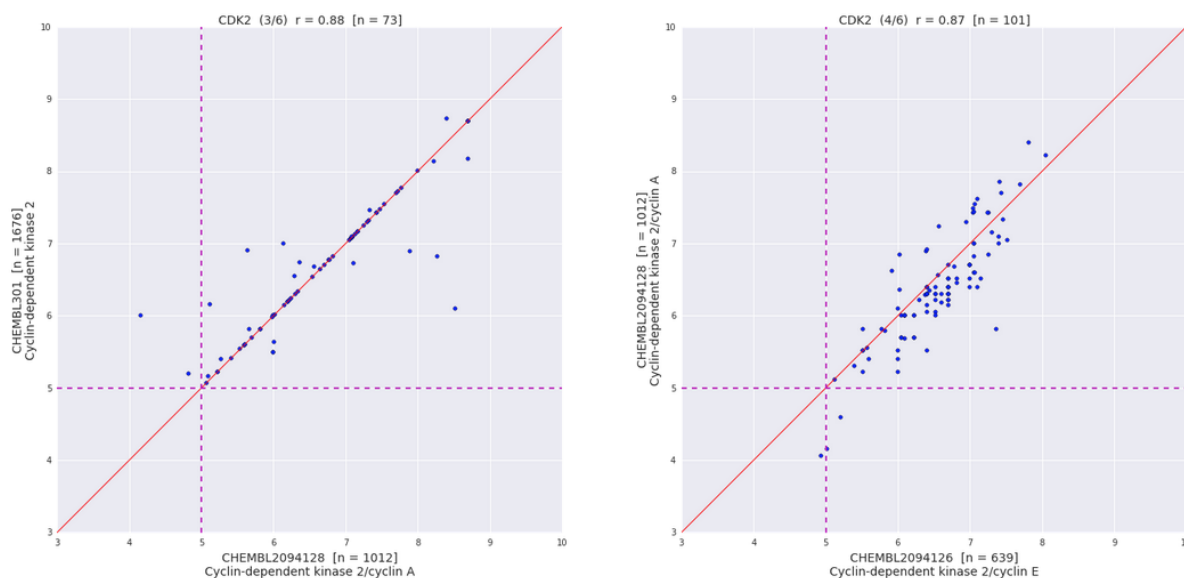
Table 1

Target ChEMBL ID	Target Type	Preferred Name	N
CHEMBL301	SINGLE PROTEIN	Cyclin-dependent kinase 2	1523
CHEMBL3038469	PROTEIN COMPLEX	CDK2/Cyclin A	88
CHEMBL3038470	PROTEIN COMPLEX	CDK2/Cyclin A1	2
CHEMBL2094128	PROTEIN COMPLEX	Cyclin-dependent kinase 2/cyclin A	865
CHEMBL2094126	PROTEIN COMPLEX	Cyclin-dependent kinase 2/cyclin E	594
CHEMBL1907605	PROTEIN COMPLEX	Cyclin-dependent kinase 2/cyclin E1	307

As mentioned in the previous section, what to do in such cases is not always obvious. Cyclins are necessary to activate CDKs, so, strictly speaking, ‘Cyclin-dependent kinase 2’ (CHEMBL301) is not a physiologically relevant target. Despite this, binding assays are sometimes used to measure ‘activities’ against such ‘inactive’ entities, and this is the source of some of the data for CHEMBL301. However, inspection of a selection of the source literature suggested that ‘Cyclin-dependent kinase 2’ or ‘CDK2’ is also often used as shorthand for a CDK2 plus cyclin complex and that this is being transcribed faithfully during the ChEMBL data-extraction process; in other words, some of the data labelled as having CDK2 alone as its target was actually generated against a more physiologically relevant complex.

To further investigate this issue, plots were generated of mean pXC50 values for pairs of ChEMBL targets associated with each gene symbol, where there were sufficient compounds in common. A selection of the plots for human CDK2 is shown in Figure 1.

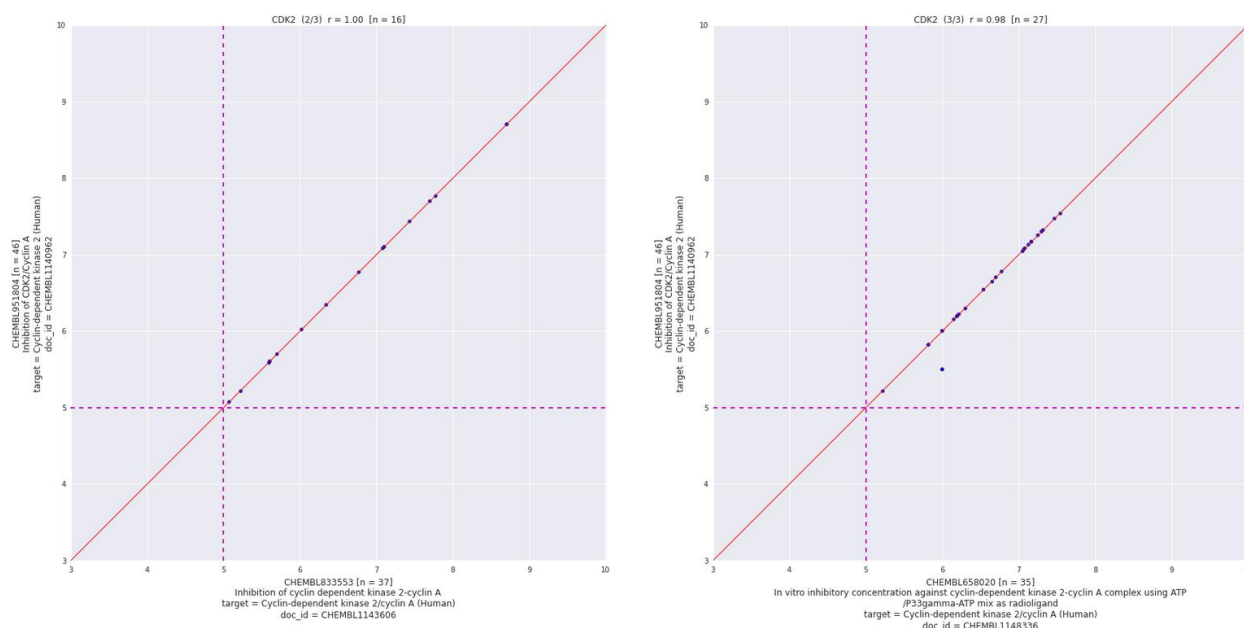
Figure 1



In general, the correlations between the data for the different CDK2/cyclin complexes are fairly good. However, some plots involving ‘Cyclin-dependent kinase 2’ (CHEMBL301) vs. a CDK2/cyclin complex (e.g. the left-hand plot in Figure 1) are problematic, showing a subset of collinear points which is unlikely to be arrived at by chance.

This issue was investigated further by plotting mean pXC50 data for pairs of ChEMBL assays. This clearly showed that three pairs of assays contain duplicated data, each of which involved an assay for the target CHEMBL301 and an assay for a CDK2/cyclin target; examples are shown in Figure 2.

Figure 2



Inspection of the source documents shows that, in each case, original data for a CDK2/cyclin complex has been replicated in another paper and mislabelled as CDK2 only. This has occurred several times for multiple targets and seems to occur most often when heterogeneous data is gathered from the literature for modelling purposes and then republished; it then becomes a single new ‘assay’ when extracted into ChEMBL. In cases where this has been spotted the details have been passed to the ChEMBL curation team.

After a round of exploratory analyses of this sort it was decided that all data for multiple ChEMBL targets corresponding to a gene symbol would be pooled. In many cases the correlations were good enough that there was no real issue; in a few cases the correlations were not so good but not to the extent that many compounds would be considered ‘active’ against one assay but inactive against another.

It should be noted that, for convenience, ‘target’ hereafter refers to a gene symbol as opposed to a ChEMBL target. The gene symbol is, of course, used to refer to its protein product, or possible the complex in which the protein is the main component. This further represents all ChEMBL targets corresponding to that protein or protein/complex.

Species

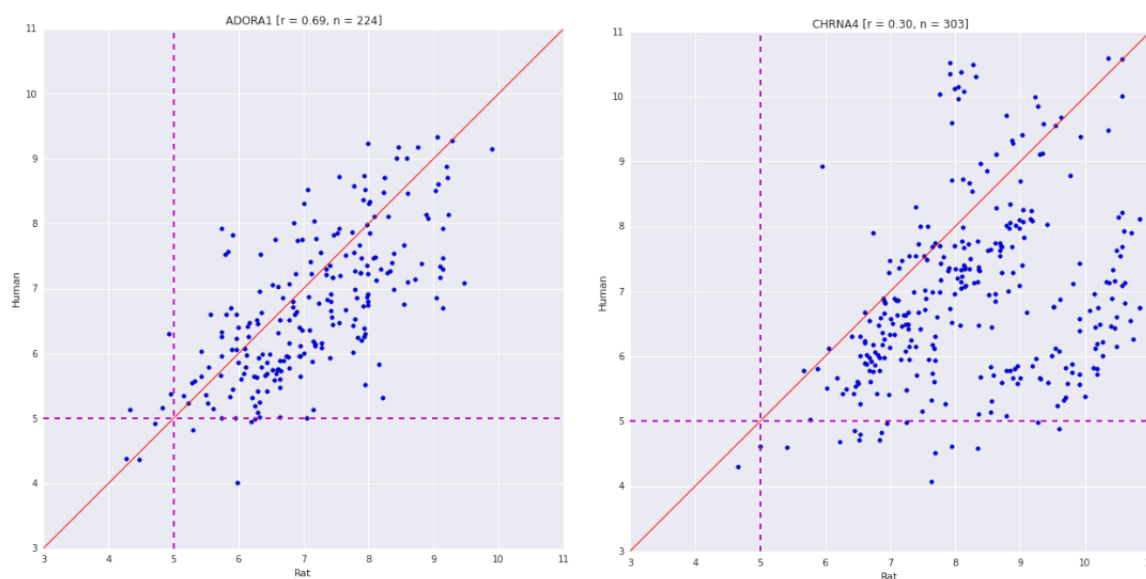
At this stage that, both human and rat data were retrieved, reflecting their relative importance in modern early-stage drug discovery. The amount of data for other species is relatively small. This is not universally true: for example, to obtain a complete picture of the SAR in a focussed study of cardiac ion channels, guinea pig, rabbit and, possibly dog data would also need to be considered. After some exploratory work, however, it seemed clear that, for this panel modelling exercise at least, adding extra species would add considerable complexity for relatively little gain.

Although the focus of HeCaToS is on modelling toxicity in humans, rat data could still have value. First is in the matter of cross-species predictivity: if activities at low-level targets are similar in the two species, then higher-level data obtained in the rat are more likely to have relevance to humans. Second, and perhaps more immediately relevant, is the possibility of data pooling across species: if the activities of compounds tested in both species are similar, then perhaps data on compounds tested only in the rat might be useful in building models intended for prediction of activities in

humans. A recent publication examined cross-species correlations across a wide variety of targets in ChEMBL [11], and found them generally concordant, although there were some notable exceptions.

The correlations were re-examined for the targets of interest, as new data could have become available in the mean time. Scatterplots of rat vs. human mean pXC50 were produced for all symbols for which there were sufficient compounds tested in common; some examples are shown in Figure 3. Generally, the plots show that cross-species correlations are reasonably good: an example of such is the Adenosine A1 Receptor (Figure 3, left). There were exceptions, however: the plot for nAChR $\alpha 4$ (Figure 3, right) suggests some systematic differences in the SAR between the species.

Figure 3



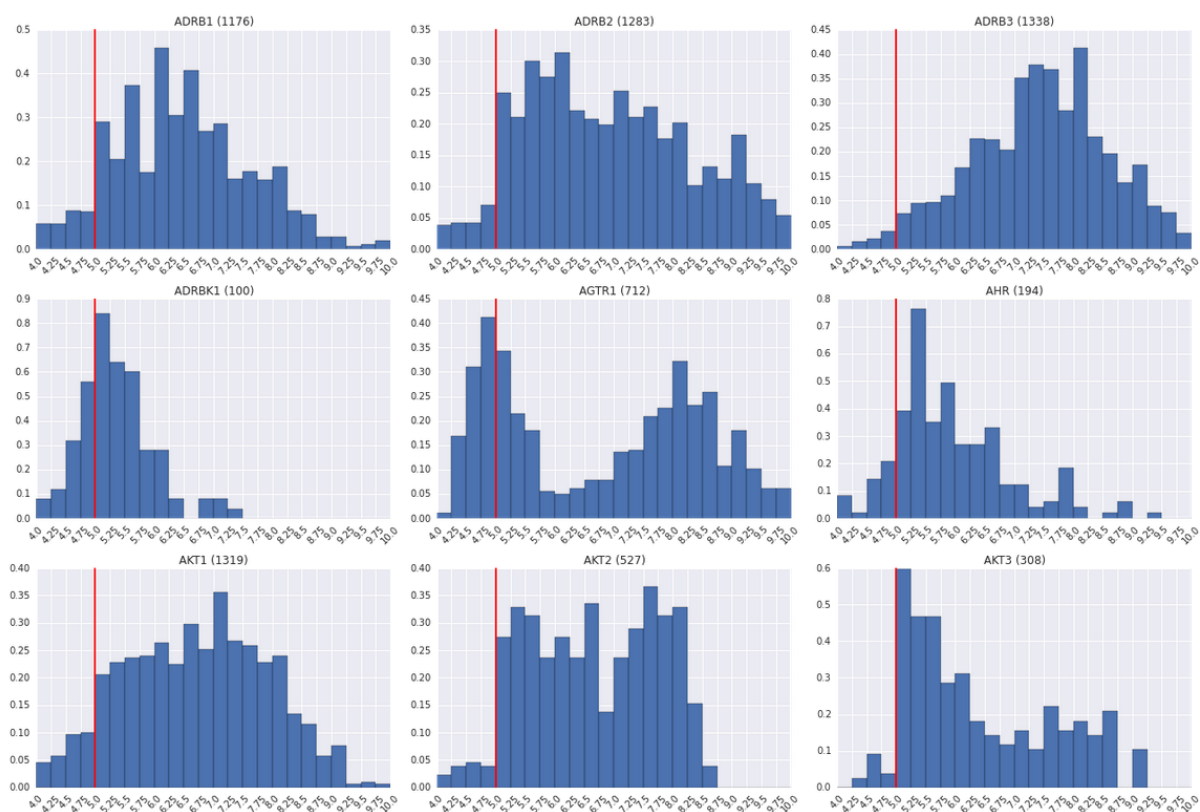
In the end, it was decided to defer pooling rat data with human data until a more thorough investigation of the effects model predictivity could be attempted. Thus, only human targets are considered going forward.

Active Compounds

At this point it is necessary to discuss what constitutes an ‘active compound’ for the purposes of this project. The exploratory analyses discussed above used mean pChEMBL values for parent compounds. The pChEMBL value [12] is the negative logarithm of the activity value (in molar units) for dose-response activity types such as IC50, XC50, EC50, AC50, K_i , K_d and Potency (with some quality filtering being applied). Thus, taking only records with a pChEMBL value is the most convenient way of obtaining the dose-response data for a target. Dose-response data is commonly preferred where available, due to the relative ease of interpretability.

Histograms of mean pChEMBL values for each target were generated, and illustrate the heterogeneity of the target activity profiles; a sample is shown in Figure 4.

Figure 4



A chart of the median activities for each target, coloured by target class, is shown in Figure 5 and the median pChEMBL value for each target class is shown in Table 2.

Figure 5

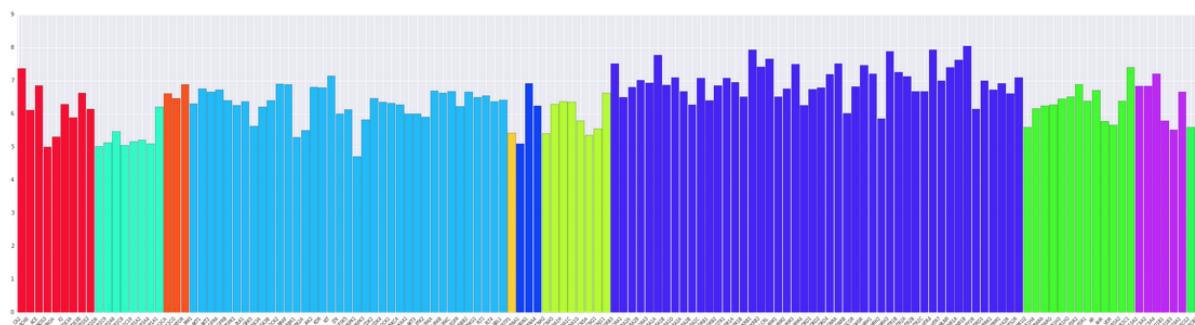


Table 2

Enzyme
Enzyme > Cytochrome P450
Enzyme > Kinase > PI3 Kinase
Enzyme > Kinase > Protein Kinase
Enzyme > Phase II
Ion channel > Ligand-gated ion channel
Ion channel > Voltage-gated ion channel
Membrane receptor > G protein-coupled receptor
Transcription factor
Transporter > Electrochemical transporter
Transporter > Primary active transporter

Target class	Median pChEMBL value for class
Enzyme > Cytochrome P450	5.1
Enzyme > Phase II	5.4
Ion channel > Voltage-gated ion channel	5.5
Transporter > Primary active transporter	5.6
Enzyme > Kinase > Protein Kinase	6.0
Ion channel > Ligand-gated ion channel	6.3
Enzyme	6.4
Transcription factor	6.6
Enzyme > Kinase > PI3 Kinase	6.6
Membrane receptor > G protein-coupled receptor	7.0
Transporter > Electrochemical transporter	7.0

It is interesting that the CYP450s have lower median activities than other major target classes. This is presumably because these are key anti-targets which drug-discovery programs optimise against, whereas the majority of, for example, GPCR activities will be derived from programs where the activity is being optimised.

In common with the ChEMBL target-prediction exercise [6] and frequent practice, a minimum pChEMBL value of 5.0 (corresponding to a IC_{50}/K_i etc. of 10 μM) is used to define an 'active' record, with values below this being defined as 'inactive'. This is a large simplification for several reasons, not least being that the various activity types are not directly comparable and the threshold is somewhat arbitrary. Given the heterogeneity of the target types and their different locations (*e.g.* cell-surface vs. intracellular) it is also unlikely that a single threshold is suitable. Nevertheless, it is considered an acceptable and necessary approximation in the circumstances.

As stated above, only active compounds are used in the MCNB modelling process. While compounds labelled as inactive are thus not used directly, they could be useful in other circumstances such as in the testing of the models. The records flagged as 'inactive' due to having a pChEMBL value below the threshold were thus supplemented by taking all other records from the assays with at least one active compound but which had no pChEMBL value.

For some targets, ChEMBL also contains other activity types such as % Inhibition. This is not used at present, as it is usually much more difficult to interpret than dose-response data. There is also some data, particularly for target classes such as transporters, where there is non-numeric data available. This is sometimes simply a text label of 'active' or 'inactive', but may include further information such as 'substrate' or 'inducer' where relevant. While this data is not currently used here, it has been included in the ChEMBL ADME SARfari [13], and its incorporation into a future iteration of this work is a possibility.

In ChEMBL, a 'parent compound' is defined by a chemical structure with any salt or solvate components removed to leave only the putative bioactive component [3]. This is suitable for most analyses, but could lead to a subtle problem in certain circumstances. The ECFP fingerprints that are to be used in the MCNB modelling encode only atom identities and connectivity: in other words, all stereoisomers for a compound will have the same fingerprint. This is a known limitation of fingerprint-based modelling techniques that is generally accepted because of their other attractive features, such as speed and simplicity.

Problems could conceivably arise if multiple stereoisomers of a compound are active against a target; for example, the presence of several active stereoisomers could lead to an overestimation of the number of distinct structures available for model building. Further, if one were used in the training set and another in the test set it could contribute to an overestimation of the prediction accuracy. It should be noted that these effects would most likely be small, and for larger datasets would be largely irrelevant. However, several of the targets of interest have relatively few data points available and might potentially be more vulnerable to these effects.

Because of this issue, it was decided to remove the possibility by using an alternative method of identifying distinct chemicals, which is the 'USMILES' [14]. This is simply a canonical SMILES generated without any stereochemical information; thus, all compounds that give the same fingerprint will share the same USMILES, and problems of the sort described above cannot occur. Using RDKit, the USMILES were generated from the SMILES for all compounds of interest, and used as the compound identifier in what follows; thus, any reference to 'compound' hereafter should be understood to refer to a distinct USMILES.

Using this definition, almost a quarter of the compound/target pairs of interest here have multiple records associated with them. In these cases a summary rule is necessary to determine whether a

compound is deemed active against a target. It was decided that if *any* pChEMBL values for a compound against a target lie above the activity threshold then that compound would be counted as an 'active' for that target.

It should be noted that this is a generous definition of what constitutes an active compound, and that other reasonable but less permissive methods were also considered. For example, all pChEMBL values for a compound could be averaged and the thresholding then applied to the mean; although for most compound/target pairs the end result is the same, there are those for which it makes a difference. After some preliminary data exploration, the more inclusive method was chosen due to the number of targets for which relatively few actives are available. There is, of course, a trade-off between increasing inclusivity and the number of 'false positives' included in the training sets and this is another topic that should be revisited in more detail later.

A small number of compounds had no structure associated with them due to ChEMBL policies. These are either biologicals, organometallic/inorganic compounds or had a molecular weight of over 1000. As none of these classes are of use in the current project the records were discarded.

In (Q)SAR modelling exercises it is recommended that structures be carefully curated beforehand [15], as a number of problems are thereby avoided. As all data used in this analysis has been drawn from ChEMBL, no such step has been performed. This is because a number of standardizations, based on those used in the FDA SRS [16], are applied to ChEMBL structures before release. For example, acidic and basic groups will be neutralized, hypervalent nitro groups converted to a charge-separated representation *etc.* A caveat is that tautomers are not canonicalised [17], thus, it is possible there will be cases where the same tautomeric moiety in different compounds will be represented differently (*e.g.* as pyridone vs. hydroxyl-pyridine). While this is unlikely to be a major issue, it could conceivably affect the results of modelling for targets where the compound numbers are particularly low. Tools are available to address this issue [18], and it will be investigated in more detail in future.

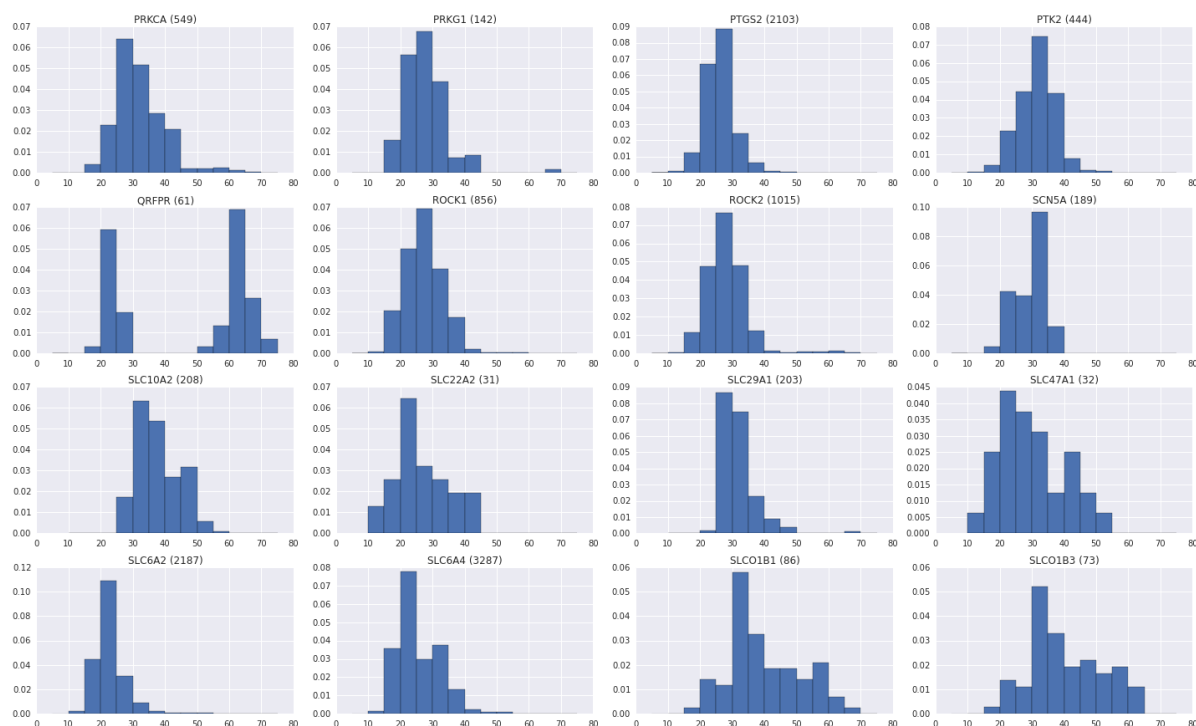
Compound Sizes

Before proceeding to the modelling, an investigation of various properties of the active compound set was undertaken. Molecular size is important, as molecules that are too large can be a problem in structure-based modelling. A large molecule will tend to have more features, and at some point this can overwhelm any 'signal' in the data, especially where the numbers of compounds is relatively small. It is thus desirable to filter out excessively large molecules before attempting to build models.

The heavy atom count was used as the key size descriptor here; while molecular weight is often used, it is probably most useful when considering distribution properties than molecular complexity. Histograms of the heavy atom count for each target show profound differences in the distributions; some example plots are shown in Figure 6. The differences are a consequence of the differences in ligand types active at the various targets. In some cases the distributions are multimodal, very obviously so in the case of such as QRFPR (Fig. 6, second row right). Inspection suggests this is due to the presence of both organic and peptidic ligands, which are unlikely to have an SAR crossover.

A cut-off of 40 heavy atoms is sometimes used and is fairly generous; however, for some targets this would still have cut the number of compounds available considerably. In the end, it was decided that a cut-off of 50 heavy atoms would be used instead, although this may need to be revisited in the future.

Figure 6

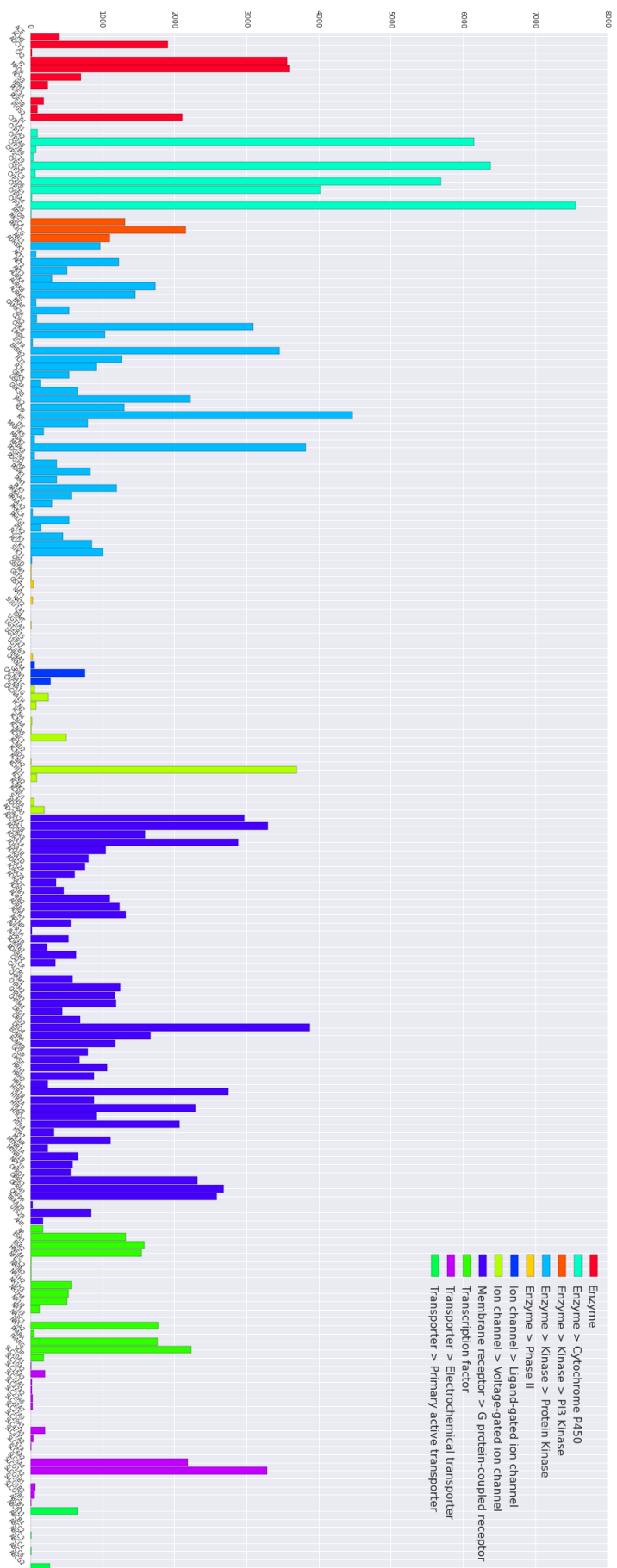


Target Classes

ChEMBL contains a hierarchical classification scheme for targets, and a simplified version was applied to the targets of interest. A plot of the count of available active compounds (filtered on size) for each target, grouped by target class is shown in Figure 7. This clearly shows how different the number of available actives can be for the different target classes. In summary:

- Enzymes that have been of interest as drug targets are well represented:
 - *e.g.* Ache, Carbonic Anhydrase, Thrombin, & Prostaglandin Synthase;
- CYP450s mainly implicated in drug metabolism are very well represented:
 - *i.e.* 1A2, 2C9, 2C19, 2D6 and 3A4;
- Kinases (protein & PI3) of therapeutic interest can be well represented:
 - *e.g.* PI3K, CDK2, EGFR & KDR;
- Phase II enzymes are very poorly represented;
- Ligand-gated ion channels are generally poorly represented:
 - the exception is therapeutic target NACHr;
- Voltage-gated ion channels are generally poorly represented:
 - the exception is key anti-target hERG;
- Various GPCRs of therapeutic interest are well represented:
 - especially adenosine, dopamine, epinephrine, histamine & opioid receptors;
- Nuclear receptors of therapeutic interest are well represented:
 - *i.e.* androgen and estrogen receptors, glucocorticoid receptor and PPARs;
- Solute carriers are generally poorly represented:
 - exceptions are the ACh and Serotonin transporters, which are therapeutic targets;
- Active transporters are poorly represented:
 - the exception is the anti-target P-gp.

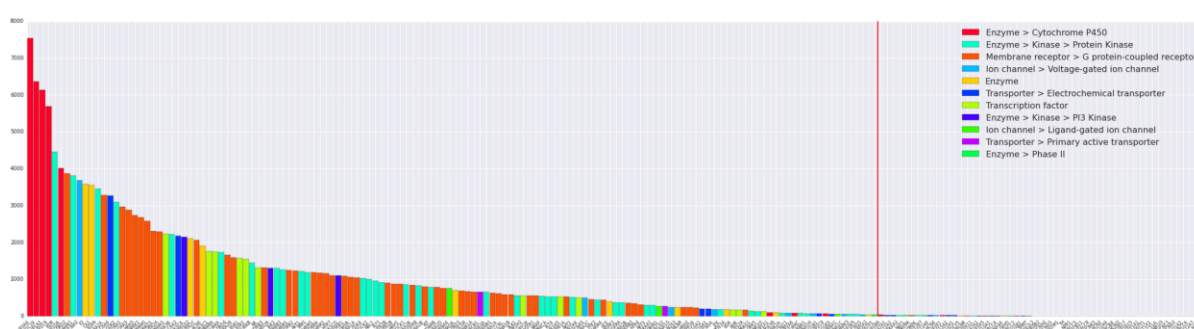
Figure 7



It is intuitively obvious that some minimum number of active compounds will be required in order to build a useful model. However, what this is is not well defined and will in fact vary depending on the structural diversity available and on the activity profile. Pragmatically, however, some threshold number must be picked in order to proceed. The ChEMBL target prediction exercise used 30 actives as the minimum training-set size, and this figure will also be used here. However, as a test set is also required, a minimum of 40 compounds per target was actually required. Thus, at a minimum, 30 compounds may be used for the model training with 10 left over for testing. Note that these figures, along with the other parameters used in this analysis, should be revisited at some point.

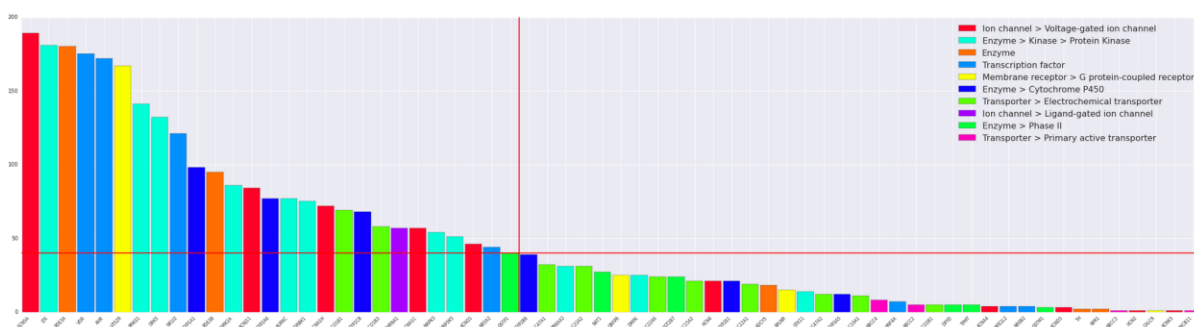
After removing those targets that did not meet the criteria for a minimum number of actives, 138 remain in the panel. The chart from Figure 7 is shown in Figure 8, this time sorted by compound count and with the threshold figure highlighted.

Figure 8



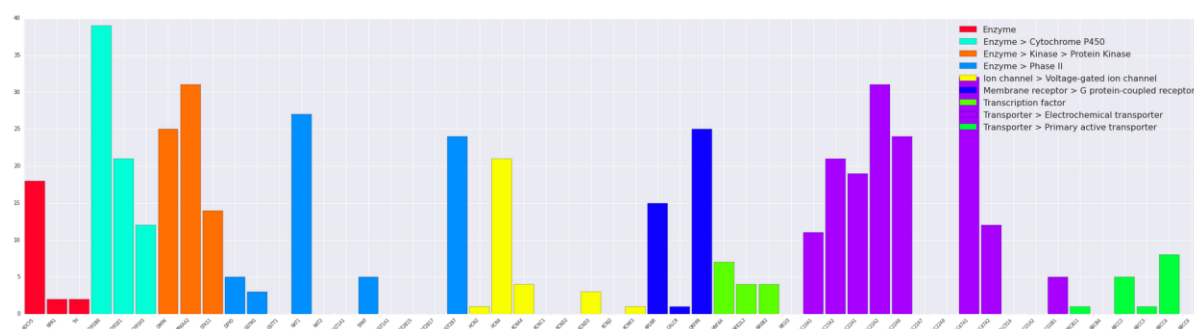
The region around the cut-off is expanded in Figure 9 to illustrate what the effect of changing the threshold might be.

Figure 9



Those targets that are eliminated using the current threshold are shown in Figure 10.

Figure 10



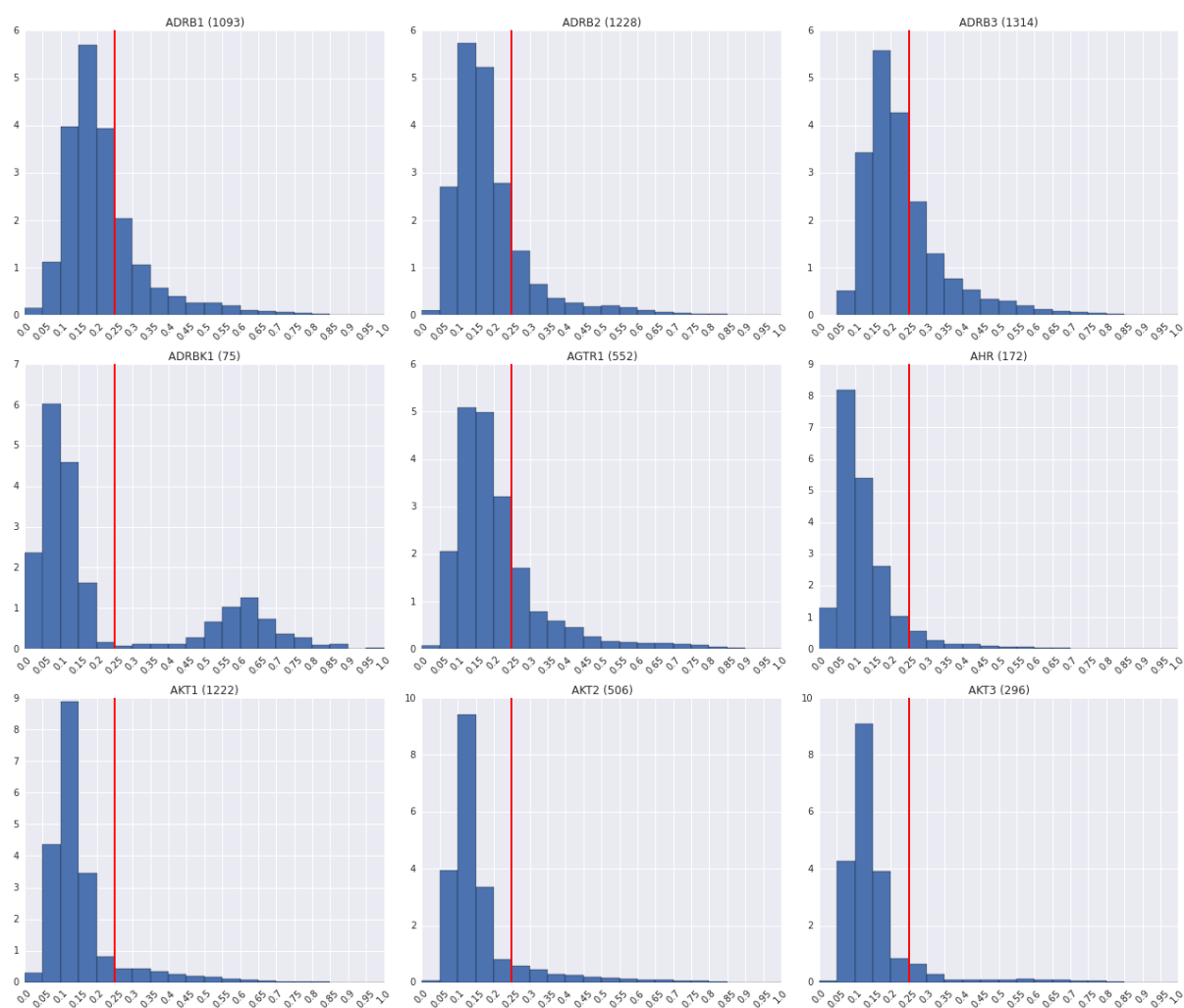
In terms of target classes, it is clear that Phase II enzymes, ion channels and transporters are particularly poorly represented in the data set. These are all important target classes for toxicological modelling and possible ways of increasing the data set for these targets will be discussed later in this document.

Compound diversity

Compound diversity is very important for any modelling technique, as unless chemical space is explored to a reasonable degree any novel compound cannot be predicted with any certainty. This is notoriously difficult to quantify, and what counts as 'sufficient' chemical diversity can not be known with certainty. Analyses carried out so far for this project have thus been essentially qualitative.

As a first step, histograms of pairwise similarities, based on ECFP fingerprints and Tanimoto distance, of the active compounds for each target were plotted; examples are shown in Figure 11.

Figure 11



The value of 0.25 is highlighted as a preliminary experiment showed that only 0.5% of pairs of compounds randomly drawn from ChEMBL has similarities about this figure; it is therefore an informal threshold below which compounds are highly unlikely to be related.

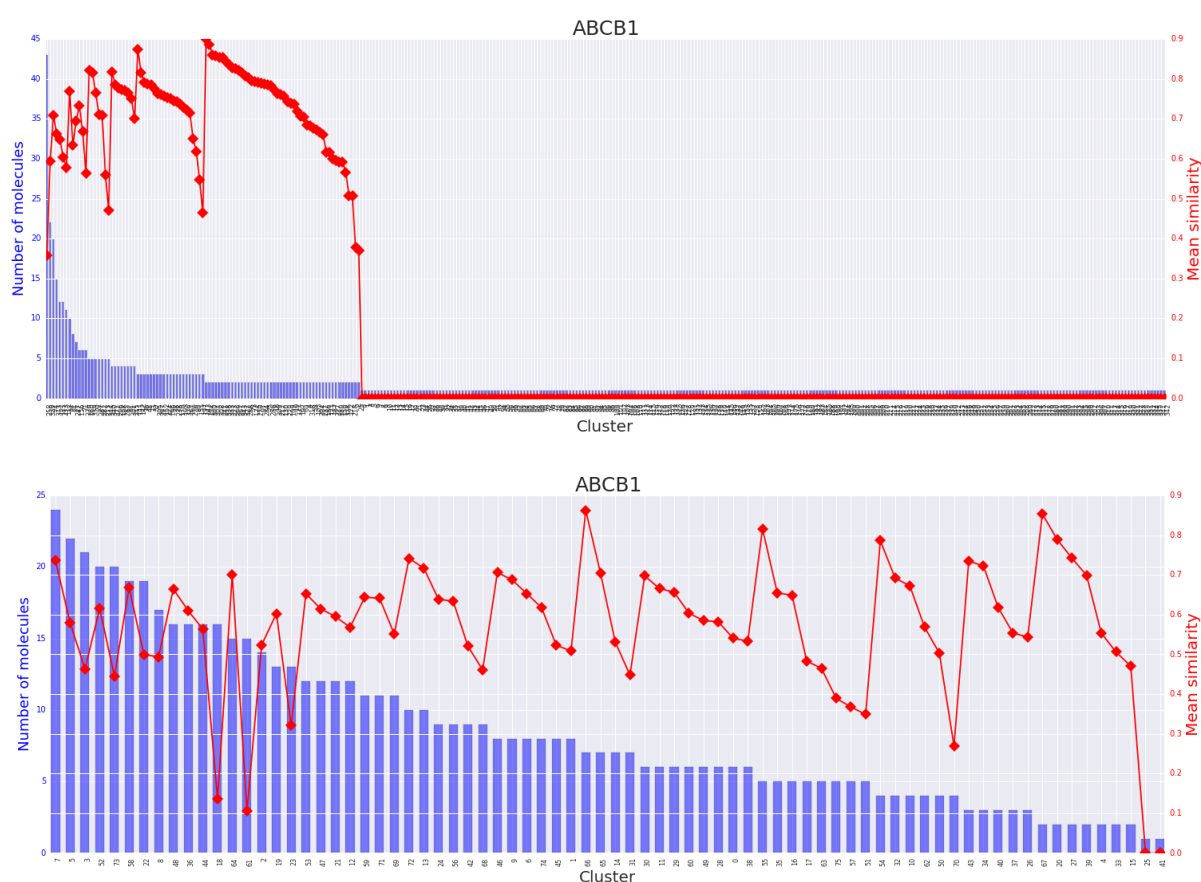
It is clear that the distributions differ for different targets. In some cases, few pairs of compounds show any meaningful similarity; in others, the distribution is more right-skewed, perhaps indicating a

higher degree of structural homogeneity. For a few targets the distribution is multimodal, which inspection confirms to be due to the presence of series of closely related compounds.

Another attempt to quantify the chemical diversity used various clustering methods to count the number of clusters for each target. The rationale here is that a larger number of clusters (relative to the number of compounds) probably indicates a more diverse dataset. After some experimentation, the most satisfactory methods seemed to be using Murko scaffolds [19] to define clusters (*i.e.* all compounds sharing a scaffold are deemed share a cluster) and Affinity Propagation [19] (*via* scikit-learn) clustering using ECFP fingerprints. These are both appealing methods that give rather different (although hopefully complementary results).

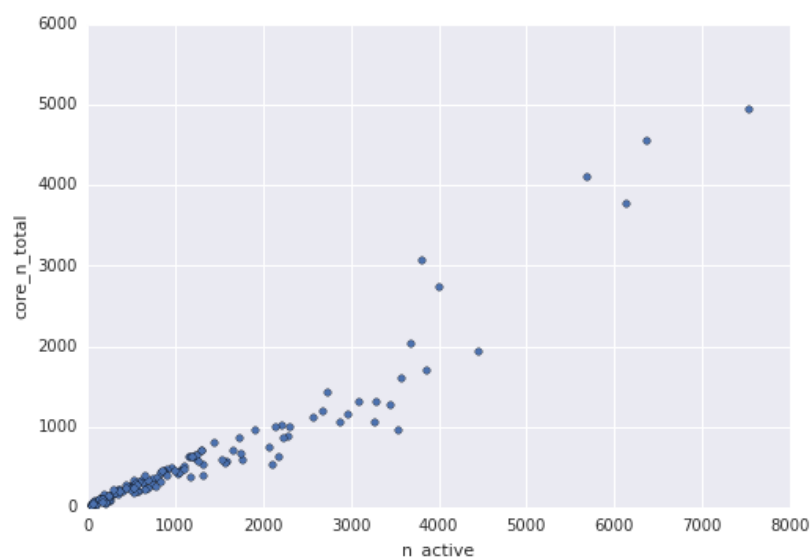
For example, the two plots in Figure 12, below, show the sizes of the clusters of active compounds for a target along with the mean intra-cluster fingerprint similarity (see below for more discussion of this). The top plot in Figure 12 uses Murko-scaffold (MS) clusters and the lower plot the Affinity-propagation (AP) clusters.

Figure 12



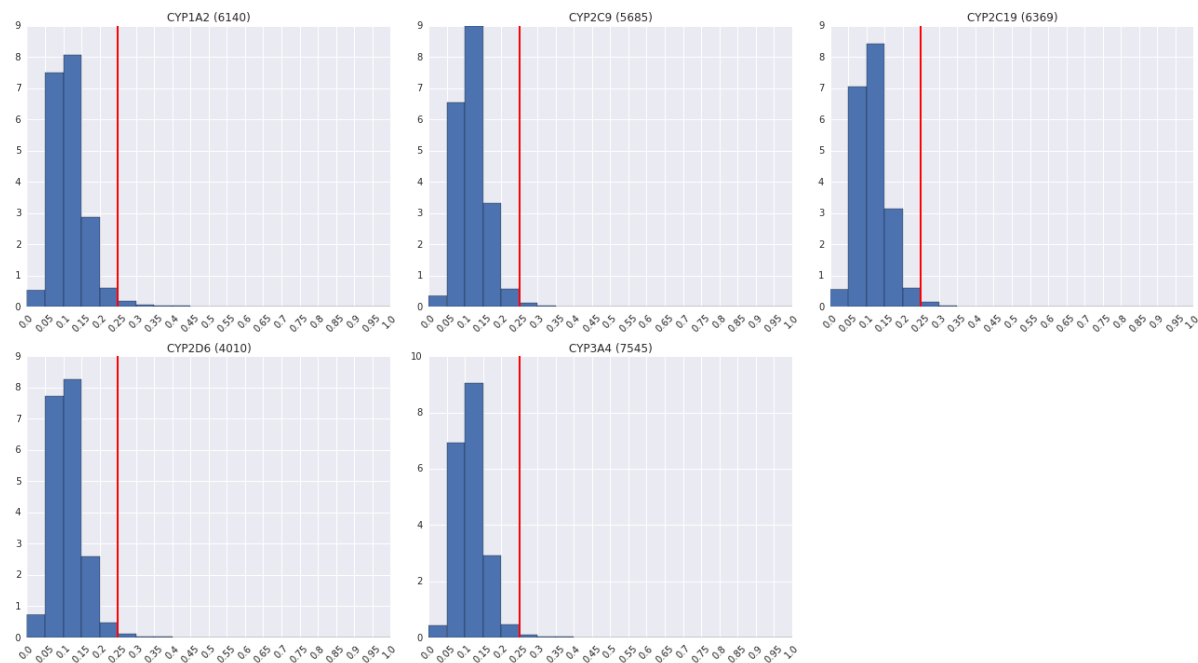
The MS method notably tends to generate many singleton clusters, whereas AP generates clusters of more balanced sizes. In general, however, the number of clusters generated by both methods tends to correlate well with the number of compounds: this is illustrated for the Murko-scaffold (*i.e.* core) clusters in Figure 13.

Figure 13



The points with the particularly large numbers of compounds in Figure 13 (*i.e.* > 4000) are mostly the CYP450s, and these seem to have more scaffold-based clusters associated with them than the overall trend would predict. This suggests the proportion of singletons is larger than normal, and hence the overall diversity higher. This is also suggested by the pairwise-similarity histograms, which, when compared with the examples shown in Figure 11, show that the CYP actives are generally less similar to each other on average than for most other target; example are shown in Figure 14.

Figure 14



This is perhaps not surprising as, as noted previously, these are key anti-targets which very many diverse projects will screen exemplar compounds in.

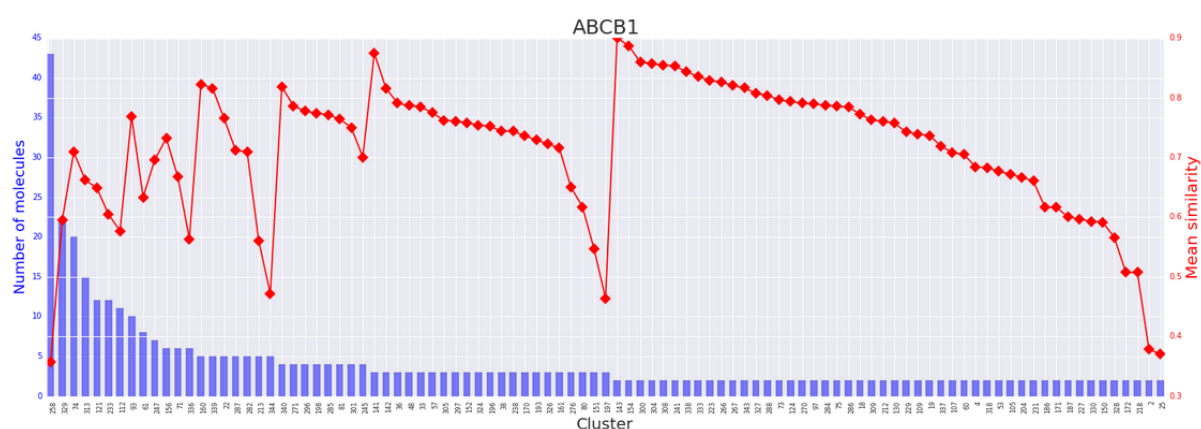
There is more analysis that remains to be done on the issue of compound diversity, and in particular on the relationship with the predictive ability of the models. The plots above mainly serve to illustrate that metrics and dashboard-style plots are straightforward to generate, and that the final

HeCaTos prediction modelling system should provide them as a matter of routine. Note that only example plots are shown here, but that the complete set is available *via* the IPython Notebooks.

One interesting avenue that is being investigated is the relationship between ‘compound series’ and whole-molecule similarity. The concept of ‘series’ is very commonly used in medicinal chemistry, although it can be remarkably hard to define. Normally, a series is a set of compounds sharing a common core substructure, also known as the ‘scaffold’ or ‘template’; the concept is approximated here by the use of Murko scaffolds to define series.

The plot shown in Figure 15 is the same as one shown at the top of Figure 12 except with the singleton clusters eliminated for clarity. Recall that a ‘cluster’ here is simply defined as all compounds with a common core (*i.e.* Murko scaffold). The ‘mean similarity’ is the average pairwise similarity for the compounds in the cluster, calculated using the ECFP4 fingerprints used throughout.

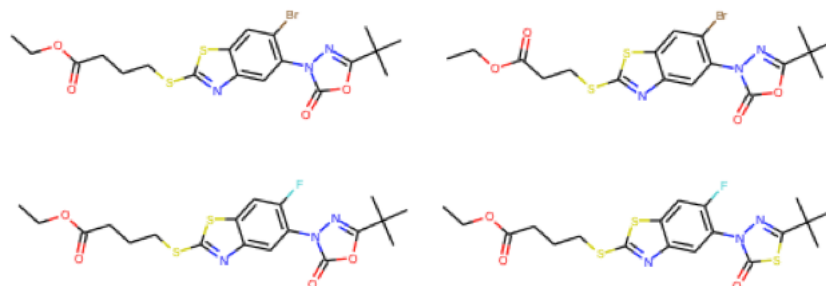
Figure 15



The variation in whole-molecule similarity among similarly sized clusters is of interest and suggests the two concepts could be complementary in describing the diversity of a dataset.

A further complication is the similarities do not always differ in an intuitive way: for example, Figure 16 shows two pairs of compounds. The top pair differ in the length of the alkyl-ester side chain and have a pairwise similarity of 0.9, while the bottom pair differ in the substitution of a sulphur for an oxygen in the oxadiazolone ring and have a pairwise similarity of only 0.74.

Figure 16



This difference in the values of the computed pairwise similarities is important, as the first pair would be classified as ‘similar’ in many applications while the second pair would not be, despite it not being obvious that the second difference is more likely to be pharmacophorically significant than the first.

Computing molecular similarity is, of course, a complex subject, made more so by the fact that it is highly context-dependent: methods that work well for one problem might not be suitable for another [20]. Nevertheless, improved ways of handling cases like these would be valuable in better quantifying the chemical diversity present in a dataset intended for modelling.

Modelling: strategy

The proposed modelling/testing protocol is to run multiple iterations of the MCNB classifier, each time dividing the available actives for each target into (different) training and test sets in a 3:1 ratio. As the MCNB technique uses those compounds in the training set that are not classed as active against a target as the inactives for that target, a parallel test set is made for each target to represent these inactives. The training and test sets are then collated and the panel model built using the training set.

The test sets, active and inactive, are then predicted using the model. The proportion of actives for a given target correctly predicted as active against that target is used as a metric of the success of the model, as is the proportion of inactives incorrectly predicted as active. Note that, given the nature on the MCNB panel model, activities against multiple targets can be predicted for a given compound; however, it is only the target(s) against which it is known to be active that are of interest here: any other predictions are discarded.

The metrics generated for each iteration of the modelling workflow are then summarised and used to judge the overall performance of the workflow.

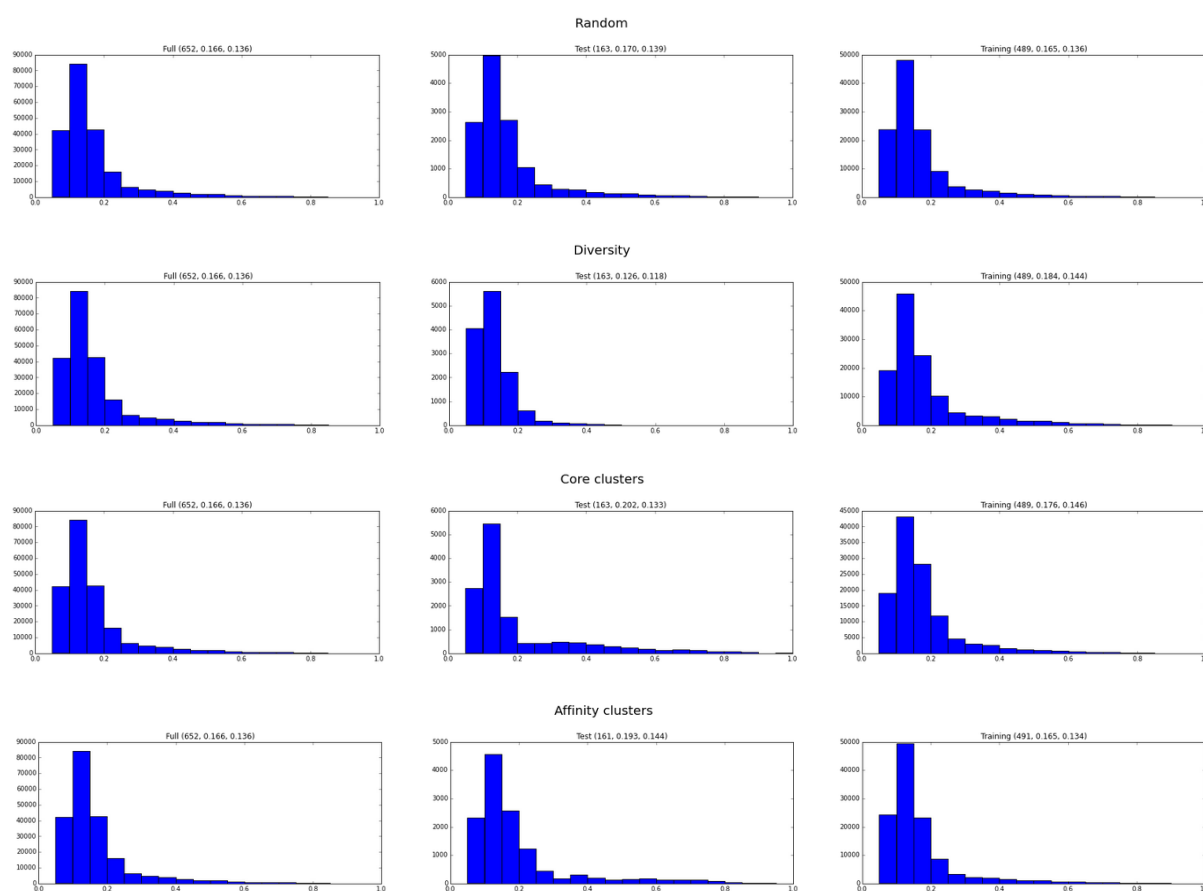
Various methods of splitting the active compounds available for a target into test and training sets were examined. Ideally, a test set would somehow represent ‘novel’ compounds, so that a realistic estimate of the model performance with genuinely novel compounds is obtained. Three methods were investigated for picking a test set: random picking, diversity picking and using picking discrete clusters.

Random picking simply means taking a random sample of the compounds. Diversity picking uses the RDKit implementation of the MaxMin algorithm [21] to choose a set of compounds as different from each other as possible. Cluster picking means selecting a random subset of clusters that together give approximately the required number of compounds; it can be done using either the Murko-scaffold based clusters or the Affinity-propagation clusters discussed above, and is designed to mimic the effect of using distinct chemical series.

These methods do pick distinctly different test sets, as can be seen by comparing the distributions of the internal pair-wise similarities of a set of compounds (the actives for typical target) and of test and training sets derived using the various methods; these are shown in Figure 17.

For the random picking (Figure 17, first row), the distributions of the full (left), test (middle) and training (right) sets are very similar. With diversity picking (second row), the distributions are noticeably different: the test set is shifted left compared to the full set, indicating lower internal similarity and thus higher diversity while the training set is commensurately shifted right, indicating it is more homogeneous than the full set. When the core clusters are used for picking (third row), both the test and training sets become more homogenous, although in slightly different ways. Picking using Affinity-propagation clusters (fourth row) gives distributions apparently similar to random picking: a finer-grained analysis of the compounds selected by both methods will be particularly interesting here.

Figure 17



Which of these partitioning methods is 'best' overall will require further work to determine. The 'core clusters' method might be expected to be the most 'realistic' and might thus be preferred where the number of (active) compounds for a target is relatively large. However, in many cases here the number of compounds is *not* large, and there are concerns as to how this method would behave in those cases, especially as the procedure needs to be iterated to generate statistics. Simple random picking has the attractive property of preserving the distribution of pairwise similarities in the test and training sets and is straightforward to implement with iteration; it is thus the obvious choice for a first, baseline, implementation and has been used as such.

Modelling: results

The plots below summarise the results obtained by running the protocol described above ten times. Note that, for the moment, default parameters are used during the model-building and prediction steps; it may therefore be possible to improve performance, for example, varying the probability threshold at which compounds are predicted to be active.

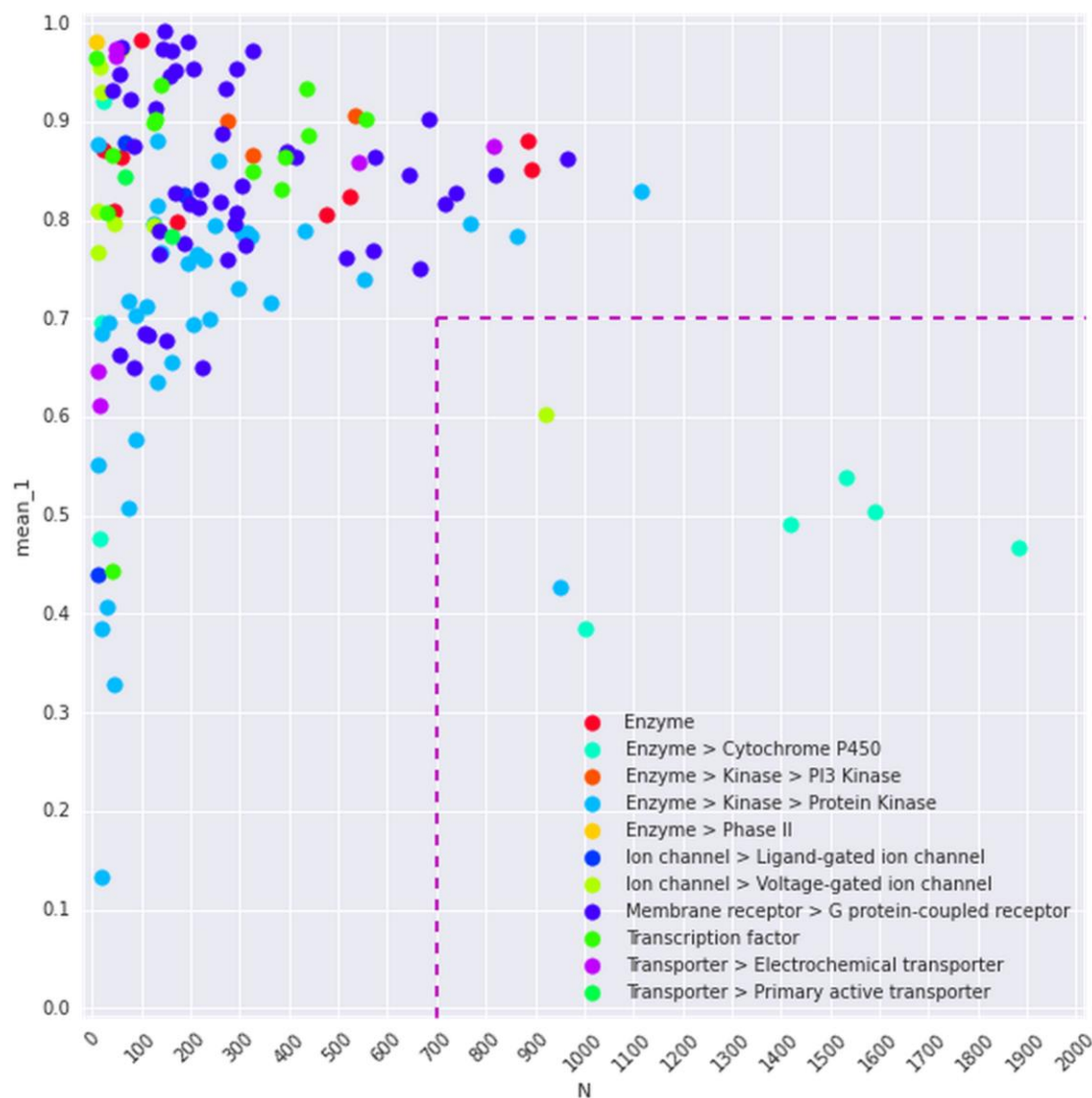
The scatterplot in Figure 18 shows the fraction of actives for each target that were correctly predicted as such (mean_1) vs. the number of actives in the test set (N).

The differences in modelling performance for different targets both between and within target classes are interesting. For example, transcription factors generally seem to perform well, GPCRs show a range of performances (which might reflect the presence of different receptor sub-types), and protein kinases perform, with exceptions, generally less well for a given dataset size.

The reasons for these differences will require further study, but likely involve differences in the activity profiles and chemical diversity for different targets and target classes. There are known to be

differences in how different classes are approached by medicinal chemists: for example, the ‘recycling’ of preveiledged scaffolds by different projects and the use of broad selectivity panels being particularly common in kinase programs.

Figure 18



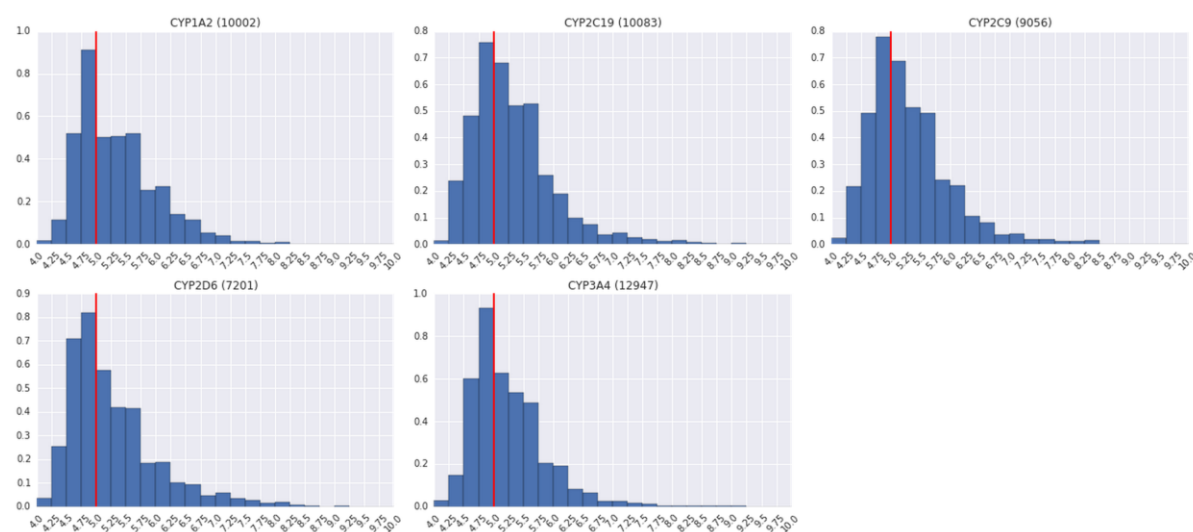
Perhaps the most immediately noticeable feature of the plot is the group of targets that are rather poorly predicted despite having a large number of actives associated with them; these are highlighted with the dashed lines in the plot and are shown in the table below.

symbol	mean_1	N	target_class
CYP3A4	0.4675	1887	Enzyme > Cytochrome P450
CYP2C19	0.5032	1593	Enzyme > Cytochrome P450
CYP1A2	0.5375	1535	Enzyme > Cytochrome P450
CYP2C9	0.4904	1422	Enzyme > Cytochrome P450
CYP2D6	0.3837	1003	Enzyme > Cytochrome P450
MAPK1	0.4262	953	Enzyme > Kinase > Protein Kinase
KCNH2	0.6025	922	Ion channel > Voltage-gated ion channel

That the CYPs should be poorly predicted is interesting but perhaps not surprising, given some results noted above. They have generally lower activities than other targets, with the modal

activities being close to (slightly below) the active/inactive threshold used here (*i.e.* 10 μ M), as shown by the activity histograms shown in Figure 19.

Figure 19

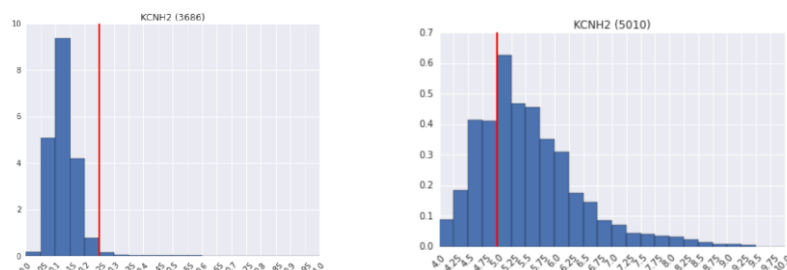


It might be that, in these cases, the active/inactive classes are particularly poorly separated and thus more difficult for the algorithm to distinguish than in other cases. In addition, as already noted, the compound sets for the CYPs are relatively heterogeneous (see Figure 14 above), which means members of the test set are less likely to resemble a members of the training set and thus be 'recognised' as active.

As discussed above, both activity profile and diversity of compounds might be a consequence of the fact that the CYPs are a key anti-target and have thus have seen a large number of diverse compounds (*i.e.* exemplars of series from many different projects) which have effectively had their activities optimised against them. These 'pure' anti-targets stand in contrast to those that are also therapeutic targets and thus have seen more compound series that have been optimised for them. This could explain the differences in activity and similarity profiles and hence modelling performance.

KCNH3, better known as hERG, is also a very important antitarget, and the same reasoning probably applies as with the CYPs. As illustrated in Figure 20, the low intra-set compound-similarity (Fig. 20 left) and activity distribution centred close to the active/inactive threshold (Fig. 20 right) are certainly similar to the CYPs.

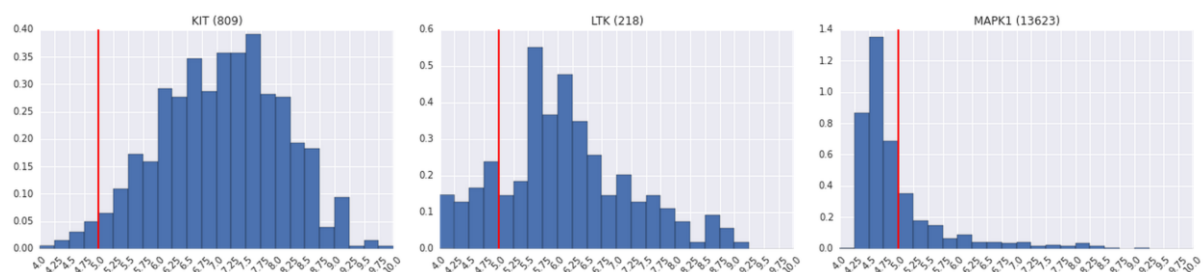
Figure 20



Where data is relatively abundant, a way to improve models might be to increase the separation between the classes by discarding 'moderately' active compounds and training only on those that are clearly active or inactive. This is not an option for many targets due to the smaller number of compounds available, but will be attempted where appropriate.

The other well-represented but poorly predicted target is MAPK1, a protein kinase. It is notable that the activity distribution of this kinase is atypical, in that it resembles those of the CYPs more than that of other kinases. In Figure 21, the activity profile for MAPK1 (Fig. 21 right) is shown alongside some more typical examples.

Figure 21



The reasons for this are not entirely clear, but the bulk of the data for MAPK1 is from a large PubChem dataset. This dataset seems to include a particularly large number of somewhat ambiguous data points.

Another notable feature is that a number of targets are predicted very poorly; whilst these have relatively few actives associated with them, by no means are all targets with a similarly small numbers of actives are badly predicted. Determining what distinguishes these categories of target will be particularly important, as it could provide insight into what it is that makes a dataset modellable.

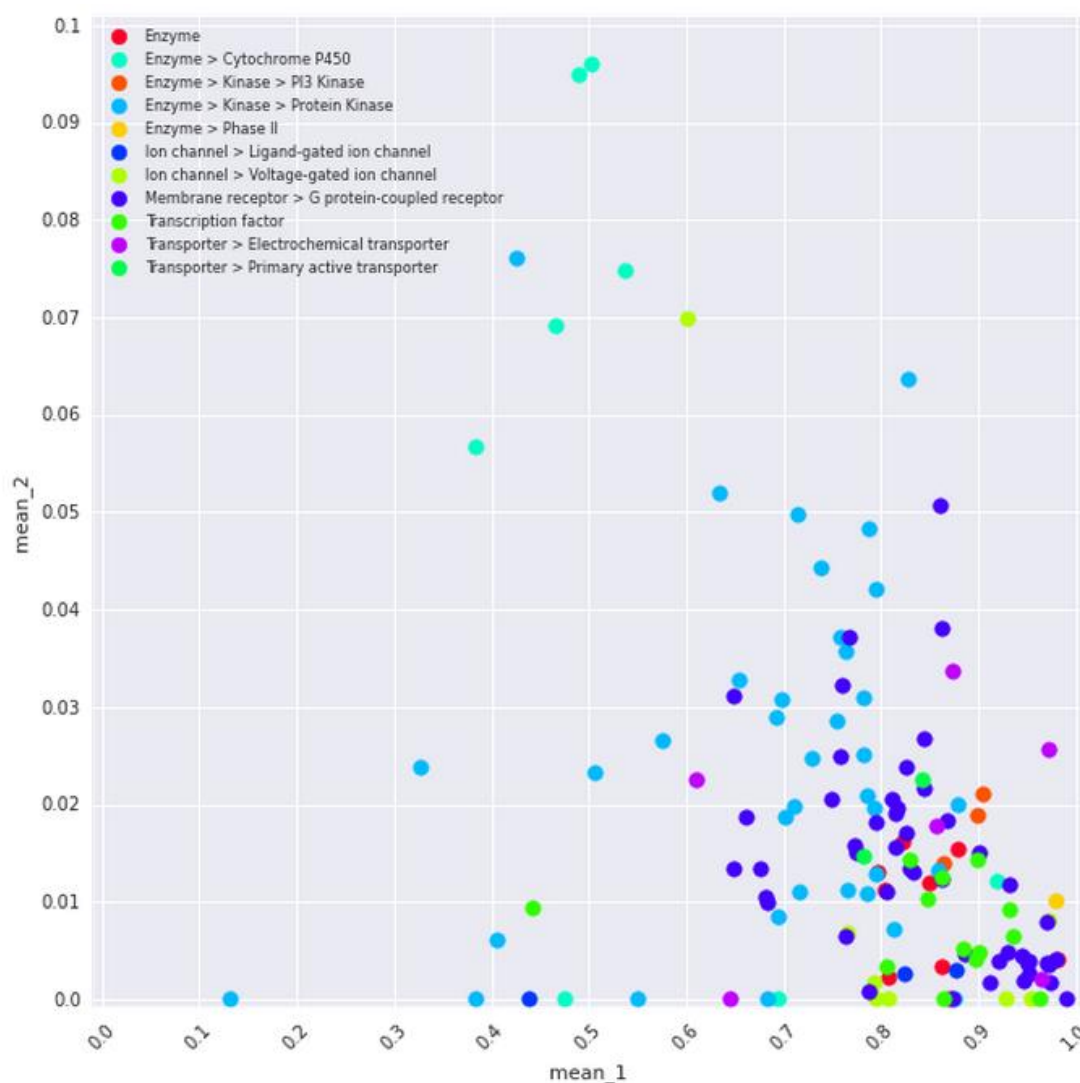
One possibility is that the active compounds for well-predicted targets are relatively homogeneous, and thus compounds in the test set tend to strongly resemble members of the training set. This would imply that the predictivity is being overestimated and that an alternative method of selecting the test set, such as using the scaffold-based clusters, might give very different results.

Another (non-mutually exclusive) possibility is that the particularly poorly predicted targets have activities centered close to the active/inactive threshold (*i.e.* 10 μM) such that the differentiation between the actives and inactives is poor. Given that literature data is being used (with the attendant inter-laboratory variation) and that different data types are being mixed (IC₅₀, K_i *etc.*), it is possible, for example, that moderately active compounds might be misclassified as inactive. If closely related compounds are classified as active, the signal in the data will be confused and poor predictivity would be expected.

As a complement to Figure 18, Figure 22 shows the mean fraction of inactives incorrectly predicted as active (*mean_2*) for each target, plotted against the correctly predicted fraction (*mean_1*).

It is notable that certain target classes, most notably transcription factors and GPCRs, tend to combine a high true-positive rate and a low false-positive rate. Again, however, the group of the CYP450s, hERG and MAPK1 stands out, having both a high false-positive rate and a low true-positive rate. The kinases seem to typically have a somewhat higher false-positive rate than other target classes; a plausible explanation for this might be the use of similar scaffolds for different kinase targets, which should be simple enough to check. Otherwise, there doesn't seem to be clear trend in the relationship between these two metrics.

Figure 22



The information for the active and inactive test sets plotted above may also be combined to give summary metrics, such as the F-score; however, this doesn't reveal any trends that the plots above do not. More sophisticated metrics, such as ROC curves, will be generated as diagnostics for future iterations of this project: however, while such analysis tools will be useful for detailed analyses of model performance, they will not change the overall picture as discussed here.

A further test has been to run a sample of 'known' inactives against the model, instead of just the 'presumed' inactives used in building the model. This shows that some models predict many 'known inactives' as active, even when the false-positive rate for 'presumed inactives' is low. For some targets, there are series of related compounds that have a range of activities spanning the active/inactive threshold; in such cases it is likely that moderately active 'inactives' structurally-related to actives will be predicted active themselves. It is hoped that closer inspection of targets like these might help improve that model-training process, perhaps by eliminating close analogues that occur in both the test and training sets.

DIFFICULTIES

Throughout this document, ways that the model-building, testing and validation process could be improved have been discussed and need not be reiterated here. These improvements will be prioritized and implemented as time allows, and will hopefully lead to incremental increases in model predictivity.

However, it should be clear that the greatest difficulty faced here is not algorithmic but rather the lack of data for many targets of interest. At every stage where compound filtering was required the most generous criteria were applied. Even so, the final panel only contains 138 out of the 215 target symbols in the original set. Furthermore, many of the issues with predictivity discussed above are most likely a consequence of limited chemical diversity.

The problem is perhaps even more acute than the overall numbers would suggest, as the lost targets are disproportionately in target classes that would be particularly interesting for advancing the state of toxicological modelling. Some classes such as aminergic GPCRs, protein kinases and nuclear receptors that contain many therapeutic targets have relatively generous amounts of data available and activity prediction (quantitative or even qualitative) is an achievable goal. However, for other important classes, such as ion channels, transporters and Phase II xenobiotic metabolising enzymes, the amount of data is relatively sparse.

As discussed in the introduction, there are several possible reasons for the lack of data for certain classes. One is that the significance of some targets has been recognized only relatively recently and thus that data generated within pharmaceutical companies has not had time to reach the literature. This would imply that the data is simply not going to be available at present, but might become available in future. Another possibility is the nature of the assays means that they are reported in journals other than those routinely covered by ChEMBL. If this were the case, useful data (albeit, perhaps, for a relatively small number of compounds) might be available but could require significant work to locate and extract.

To what extent either of these hypotheses is true is not known and some effort will be required to clarify the situation. To this end, some preliminary work has been undertaken at EMBL-EBI on strategies for finding relevant data in the literature in a 'high-throughput' manner. This will be discussed in more detail elsewhere, but briefly involves programmatically:

- Searching MedLine using selected keywords;
- Ranking of abstracts using a ChEMBL-likeness score [22];
- Identification of whether full-text is available (*e.g. via PubMedCentral or campus library*);
- Marking up abstracts for inspection (highlighting keywords, chemical terms *etc.*).

The inspection of the abstracts and the downloading, inspection and selection of documents still need to be performed manually, although work is on going to see to what extent those can be automated. For example, there are tools available to extract chemical structures from PDFs, and these are being evaluated. However, the final step, where useable data is extracted from the selected documents, remains the most time-consuming as this must still be done manually. While ChEMBL is based around data extracted from the literature by a Contract Research Organisation, to what extent it would be feasible to use this facility on behalf of a project such a HeCaToS is not clear at present.

In the short term, focussed literature mining for a small number of key targets is perhaps viable, and the L-Type Calcium Channel is being used as a test case as it of particular interest to HeCaToS. In addition, it is possible that contacts within the pharmaceutical industry might be willing and able to make public some relevant datasets, and this possibility is being explored.

REFERENCES

1. Allen, T.E., et al., *Defining molecular initiating events in the adverse outcome pathway framework for risk assessment*. Chem Res Toxicol, 2014. **27**(12): p. 2100-12.
2. Vinken, M., *The adverse outcome pathway concept: a pragmatic tool in toxicology*. Toxicology, 2013. **312**: p. 158-65.
3. Bento, A.P., et al., *The ChEMBL bioactivity database: an update*. Nucleic Acids Res, 2014. **42**(Database issue): p. D1083-90.
4. Nidhi, et al., *Prediction of biological targets for compounds using multiple-category Bayesian models trained on chemogenomics databases*. J Chem Inf Model, 2006. **46**(3): p. 1124-33.
5. Rogers, D. and M. Hahn, *Extended-Connectivity Fingerprints*. Journal of Chemical Information and Modeling, 2010. **50**(5): p. 742-754.
6. Papadatos, G. *Ligand-based target predictions in ChEMBL*. 2014; Available from: <http://chembl.blogspot.co.uk/2014/04/ligand-based-target-predictions-in.html>.
7. Croft, D., et al., *The Reactome pathway knowledgebase*. Nucleic acids research, 2014. **42**(D1): p. D472-D477.
8. PyData. *PyData: Python Data Tools*. 2015; Available from: <http://pydata.org/downloads/>.
9. Landrum, G. *RDKit: Open-Source Cheminformatics Software*. 2015; Available from: <http://rdkit.org/>.
10. HUGO. *HGNC: The HUGO Gene Nomenclature Committee*. 2015; Available from: <http://www.genenames.org/>.
11. Kruger, F.A. and J.P. Overington, *Global analysis of small molecule binding to related protein targets*. PLoS Comput Biol, 2012. **8**(1): p. e1002333.
12. Papadatos, G., et al., *Activity, assay and target data curation and quality in the ChEMBL database*. J Comput Aided Mol Des, 2015.
13. Davies, M., et al., *ADME SARfari: Comparative Genomics of Drug Metabolising Systems*. Bioinformatics, 2015.
14. Weininger, D. *SMILES Tutorial: Related languages*. 1998; Available from: <http://www.daylight.com/meetings/summerschool98/course/dave/smiles-relatives.html>.
15. Fourches, D., E. Muratov, and A. Tropsha, *Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research*. J Chem Inf Model, 2010. **50**(7): p. 1189-204.
16. Administration, U.S.F.a.D., *Food and Drug Administration Substance Registration System Standard Operating Procedure Version 5c*. 2007.
17. Hersey, A., et al., *Chemical databases: curation or integration by user-defined equivalence?* Drug Discovery Today: Technologies, 2015. **14**: p. 17-24.
18. Atkinson, F.L. *standardiser*. 2014; Available from: <https://wwwdev.ebi.ac.uk/chembl/extra/francis/standardiser/>.
19. Bemis, G.W. and M.A. Murcko, *The Properties of Known Drugs. 1. Molecular Frameworks*. Journal of Medicinal Chemistry, 1996. **39**(15): p. 2887-2893.
20. Maggiora, G., et al., *Molecular Similarity in Medicinal Chemistry*. Journal of Medicinal Chemistry, 2014. **57**(8): p. 3186-3204.
21. Ashton, M., et al., *Identification of Diverse Database Subsets using Property-Based and Fragment-Based Molecular Descriptions*. Quantitative Structure-Activity Relationships, 2002. **21**(6): p. 598-604.
22. Papadatos, G., et al., *A document classifier for medicinal chemistry publications trained on the ChEMBL corpus*. Journal of cheminformatics, 2014. **6**(1): p. 40.

D1.5 APPENDIX A

This document comprises Appendix A of the HeCaTos Deliverable report D1.5 'Package of Predictive Models'. It is effectively an updated version of EMBL-EBI HeCaToS deliverable D9.2 'Report on reference set of toxicology associated targets', with some extra targets, discussion and references added.

An Excel Workbook containing the tables of targets discussed here is included as D1.5 Appendix B.

Toxicology-associated targets

Introduction

The aim of this report is to present a reference set of toxicology-associated targets. The concept of ‘target’ in pharmacology and toxicology can be used to mean different things; for example, in attempts to model drug-induced liver injury (DILI), the organ itself is sometimes considered as the target of a chemical [1]. For a variety of reasons, including the multiple mechanisms of hepatotoxicity and many possible confounding factors [2, 3], this whole-organ approach has had limited success.

In this report, we will generally consider a ‘target’ to be a specific biomolecule, *i.e.* a protein or protein complex of defined stoichiometry with which xenobiotics might interact to produce an adverse response; these are often referred to as anti-targets or off-targets in order to distinguish them from therapeutic targets [4, 5]. Note that a biomolecule may be considered a therapeutic target or anti-target depending on context; it can be beneficial to engage a target in a particular disease state, but harmful in another. A number of databases of toxicity-associated targets already exist [6, 7], which, while not exhaustive, do nevertheless contain useful information.

Restricting the definition to molecular-level targets is appropriate for several reasons. For example, such an interaction could be the molecular initiating event (MIE) [8] of an adverse outcome pathway (AOP) [9], a concept increasingly being used to formalise thinking about adverse responses to xenobiotics. In addition, such molecular-level data could be used to build computational models [10] underpinning a multi-scale approach to toxicity prediction [11, 12], an idea and modelling strategy which itself meshes well with the AOP concept [13].

For certain classes of targets such as GPCRs, nuclear receptors or ligand-gated ion channels, there can be multiple modes of engagement, such as agonism, antagonism and perhaps inverse agonism [14]. Toxic effects might be caused by one or other mode of engagement and, when annotating anti-targets with associated toxicities, the mode should be specified whenever possible. However, this information is not always known, such as where the only data is from a binding assay (yielding *e.g.* a K_d end point), and this can complicate the interpretation and use of the data considerably.

Although MIEs often involve binding to the receptors and enzymes on which we shall focus here, they may also involve less specific events such as protein alkylation [15]. In addition, a toxic insult, like a therapeutic effect, may require engagement of more than one target and it might be the activity *profile* that is important [16, 17]. In other words, at least in some cases, engagement of a combination of individually innocuous targets might be required to induce a toxic effect. This would make associating targets with adverse drug reactions (ADR) more difficult and also complicate modelling strategies. However, there is also evidence that side effects tend to be mediated by interaction of drugs with individual proteins [18].

Given the strategic aims of the HeCaToS project, the focus will be on anti-targets of relevance to hepatotoxicity and cardiovascular toxicity and on the machinery controlling drug disposition. In compiling the report, the decision was made to include only those targets that have some level of validation available. There have been various publications describing attempts to link molecular targets to drug adverse events or side effects statistically [18-21], but, while these *potential* novel anti-targets are interesting, they do not tend to come with any independent verification or with a clear mechanistic rationale. Such targets will thus be kept in mind for future data-gathering and/or modelling purposes, but will not be considered further at this time.

A pragmatic method of identifying anti-targets with at least some level of validation is to take those that are used for *in vitro* profiling in a drug discovery setting [5, 22], where they are either published by pharmaceutical companies [23, 24] or appear in the assay catalogues of contract research

organisations (CRO) used by these companies [25, 26]. This should give a fairly conservative set of anti-targets, which could then perhaps be augmented with newer or more speculative examples from the literature.

One issue with this approach is that the targets in these lists are not always annotated with the organ or tissue in which they exert a toxic effect, so it might not be apparent which contribute to cardiovascular or hepatic toxicity specifically. This could potentially be addressed to some extent by automatic annotation using a tissue expression database [27, 28], although analysis of the literature for each target individually would be required for confidence in the conclusions. Furthermore, the reasons for inclusion (*i.e.* the weight of evidence, a mechanistic rationale *etc.*) are not always given and, again, this could only really be addressed by consulting the literature.

The effects of xenobiotics on mitochondria are very important for understanding both hepatotoxicity [29, 30] and cardiotoxicity [31, 32] and modelling of mitochondrial toxicity is considered an important part of EMBL-EBI's contribution to HeCaToS WP1. Mitochondrial biochemistry is well described by reaction/pathway databases such as Reactome [33]; this is valuable as pathways can be used as an organising framework for modelling activities.

Thinking in terms of pathways and systems, in mitochondria and beyond, provides deeper insight into mechanisms of toxicity and is clearly compatible with the AOP and multi-scale concepts discussed above. Data for mitochondrial targets could also be of use for the dynamic modelling [34] strategies being pursued in HeCaToS WP2. This in turn could inform WP1 activities, for example by identifying those components of pathways that are most likely to disrupt proper cell functioning if inhibited and which should therefore be prioritised for further investigation.

The xenobiotic metabolising enzymes (XME) and transporters involved in drug disposition are another special class of anti-target [35]. Although they can be involved in direct toxicity [3], interactions with these entities are most likely to cause problems in indirect ways, such as by generating reactive metabolites or by altering the distribution or metabolism of co-administered species and causing so-called drug-drug interactions (DDI) [36]. These entities are thus of interest for toxicity prediction, even where they do not quite fit the definition of anti-target given above.

When considering the potential for DDIs, organs other than those that are the focus of HeCaToS must also be considered. XMEs and transporters are present in many tissues such as the GI tract and the kidneys [37, 38], and inhibition or induction of any of these by a drug could have an impact on the concentration of other species.

Although the focus of this report will be on molecular targets, this does not mean that data gathering should be restricted to such assays. Data for assays conducted on other levels will also be valuable for multi-scale modelling, both for validating bottom-up models and for enabling complementary 'middle-out' approaches [39]. An example would be the use of cell-based assays for measuring drug-induced mitochondrial dysfunction [40], or high-content screening assays for DILI [41]. Even compendia of hepatotoxic drugs [42], while not necessarily attractive for modelling purposes, could be useful for testing hypothesis generated from lower level models.

Although the focus is on toxicants that exert their effect *via* molecular targets, there are some of interest to this project that do not exert their effect by direct interactions with biomolecules, but rather *via* intrinsic chemical properties: examples would be mitochondrial uncouplers[43] and compounds capable of redox-cycling [44].

Cardiovascular Targets

In some cases, as with hERG and arrhythmias, the link between an anti-target and the associated toxicity is relatively well (although not fully) understood [45]. Further, in this case, the anti-target is clearly localised to cardiac tissue. In some cases, however, the linkage is less direct and possibly multifactorial [46, 47]. An example would be the effects of NSAIDs, where COX2 inhibition reduces prostaglandin synthesis, leading to vasoconstriction and, *via* effects on the kidneys, to water retention; together, these effects can lead to heart failure in vulnerable populations [47]. Such complexities must always be borne in mind when attempting to link data for molecular anti-targets to organ-level (or higher) effects.

A recent perspective on the use of *in vitro* profiling in the drug discovery process gave a list of targets recommended by representatives of multiple pharmaceutical companies as a core panel for early assessment of possible safety-related liabilities [24]. These were annotated with the organ(s) primarily affected and a list of pathological effects (with references), broken down by interaction type. Of the 44 listed, 30 have the cardiovascular system (CVS) as one of the organs affected: these are shown in Table 1. The effects are broken down by agonism/activation vs. antagonism/inhibition, the importance of which was noted above, and references are provided for each target. Both because of the provenance and the level of annotation, this would seem to be an excellent starting set of cardiovascular targets.

Note that, in some cases, the effects listed do not seem to include any obviously cardiovascular in nature; for examples, see the μ -opioid receptor and the serotonin transporter. However, a brief inspection of the literature suggests there is evidence of μ -opioid receptors affecting the CVS [48]. Further, given there are several serotonin receptors included, that the associated transporter should be an anti-target would certainly seem plausible. These cases illustrate how further research and/or mechanistic thinking might be necessary to understand the rationale for the inclusion of anti-targets, even in a list as well-annotated as this one.

It is explicitly stated that this consensus list comprises a *minimal* set of assays, and that the companies involved all screen other targets in addition. For example, an earlier, but still widely cited, review of the same topic included a list of cardiovascular targets screened during the discovery phase at Novartis [23]; these are presented in Table 2, with those also appearing in the consensus list in Table 1 highlighted.

It is interesting that many of the extra targets belong to protein families already seen (*e.g.* various GPCR subtypes [49, 50] or ion channels [51]) or are part of common pathways or systems (*e.g.* the ATP-sensitive K^+ channel has a role in the cardiac action potential – see below). Pursuing these types of relationship (*i.e.* homology or mechanism) in the literature would be a rational way of expanding the list of anti-targets, if that should prove desirable; this topic is discussed at more length below.

Note that, although some information on possible ADRs is provided in this second list, it is not split out by interaction type (*e.g.* agonism vs. antagonism). This is information that would be vital in linking target interactions to phenotypes, and thus might require further literature research.

The Novartis group also later published another version of their cardiac safety panel, which includes some targets not mentioned in the earlier version [5]. The list also separates out the effects of agonists and antagonists where relevant. There are also some further targets mentioned in a footnote to the main table and described as 'Additional targets in the selectivity panel'. This table is not included here, but is available in the accompanying workbook (as Table 2.2).

Table 1: Cardiovascular targets taken from Table 1 in reference [24]. Note that there are references for each entry in the original.

Target	Gene	Organ(s)	Effects Agonism or activation	Effects Antagonism or inhibition
Adenosine receptor A2A	ADORA2A	CVS, CNS	Coronary vasodilation; decrease in BP and reflex; increase in HR; decrease in platelet aggregation and leukocyte activation; decrease in locomotor activity; sleep induction	Potential for stimulation of platelet aggregation; increase in BP; nervousness (tremors, agitation); arousal; insomnia
α_{1A} -adrenergic receptor	ADRA1A	CVS, GI, CNS	Smooth muscle contraction; increase in BP; cardiac positive inotropy; potential for arrhythmia; mydriasis; decrease in insulin release	Decrease in smooth muscle tone; orthostatic hypotension and increase in HR; dizziness; impact on various aspects of sexual function
α_{2A} -adrenergic receptor	ADRA2A	CVS, CNS	Decrease in noradrenaline release and sympathetic neurotransmission; decrease in BP; decrease in HR; mydriasis; sedation	Increase in GI motility; increase in insulin secretion
β_1 -adrenergic receptor	ADRB1	CVS, GI	Increase in HR; increase in cardiac contractility; electrolyte disturbances; increase in renin release; relaxation of colon and oesophagus; lipolysis	Decrease in BP; decrease in HR; decrease in CO
β_2 -adrenergic receptor	ADRB2	Pulmonary, CVS	Increase in HR; bronchodilation; peripheral vasodilation and skeletal muscle tremor; increase in glycogenolysis and glucagon release	Decrease in BP
Dopamine receptor D ₁	DRD1	CVS, CNS	Vascular relaxation; decrease in BP; headaches; dizziness; nausea; natriuresis; abuse potential	Dyskinesia; parkinsonian symptoms (tremors); anti-emetic effects; depression; anxiety; suicidal intent
Dopamine receptor D ₂	DRD2	CVS, CNS, endocrine	Decrease in HR; syncope; hallucinations; confusion; drowsiness; increase in sodium excretion; emesis; decrease in pituitary hormone secretions	Orthostatic hypotension; drowsiness; increase in GI motility
Endothelin receptor A	EDNRA	CVS, development	Increase in BP; aldosterone secretion; osteoblast proliferation	Teratogenicity
Histamine H1 receptor	HRH1	CVS, immune	Decrease in BP; allergic responses of flare, flush and wheal; bronchoconstriction	Sedation; decrease in allergic responses; increase in body weight
Histamine H2 receptor	HRH2	GI, CVS	Increase in gastric acid secretion; emesis; positive inotropy	decrease in gastric acid secretion
δ -type opioid receptor	OPRD1	CNS, CVS	Analgesia; dysphoria; psychomimetic effects; cardiovascular effects; convulsion	increase in BP; increase in cardiac contractility
κ -type opioid receptor	OPRK1	GI, CNS, CVS	decrease in GI motility; increase in urinary output; sedation and dysphoria; confusion; dizziness; decrease in locomotion; tachycardia	Insufficient information
μ -type opioid receptor	OPRM1	CNS, GI, CVS	Sedation; decrease in GI motility; pupil constriction; abuse liability; respiratory depression; miosis; hypothermia	increase in GI motility; dyspepsia; flatulence
Muscarinic acetylcholine receptor M ₁	CHRM1	CNS, GI, CVS	Proconvulsant; increase in gastric acid secretion; hypertension; tachycardia; hyperthermia	decrease in cognitive function; decrease in gastric acid secretion; blurred vision
Muscarinic acetylcholine receptor M ₂	CHRM2	CVS	decrease in HR; reflex; increase in BP; negative chronotropy and inotropy; decrease in cardiac conduction (PR interval); decrease in cardiac action potential duration	Tachycardia; bronchoconstriction; tremors
5-HT _{1B}	HTR1B	CVS, CNS	Cerebral and coronary artery vasoconstriction; increase in BP	increase in aggression
5-HT _{2A}	HTR2A	CVS, CNS	Smooth muscle contraction; platelet aggregation; potential memory impairments; hallucinations; schizophrenia; serotonin syndrome	Insufficient information
5-HT _{2B}	HTR2B	CVS, pulmonary, development	Potential cardiac valvulopathy; pulmonary hypertension	Possible cardiac effects, especially during embryonic development

Vasopressin V _{1A} receptor	AVPR1A	Renal, CVS	Water retention in body; increase in BP; decrease in HR; myocardial fibrosis; cardiac hypertrophy; hyponatraemia	Insufficient information
Acetylcholine receptor subunit α 1 or α 4	CHRNA1, CHRNA4	CNS, CVS, GI, pulmonary	Paralysis; analgesia; increase in HR; palpitations; nausea; abuse potential	Muscle relaxation; constipation; apnoea; decrease in BP; decrease in HR
Voltage-gated calcium channel subunit α Cav1.2	CACNA1C	CVS	Insufficient information	Vascular relaxation; decrease in BP; decrease in PR interval; possible shortening of QT interval of ECG
Potassium voltage-gated channel, subfamily H member 2 (hERG)	KCNH2	CVS	Insufficient information	Prolongation of QT interval of ECG
Potassium voltage-gated channel KQT-like member 1 and minimal potassium channel MinK	KCNQ1 & KCNE1	CVS	Atrial fibrillation	Long QT syndrome; potential hearing impairment, deafness and GI symptoms
Voltage-gated sodium channel subunit α Nav1.5	SCN5A	CVS	Insufficient information	Slowed cardiac conduction; prolonged QRS interval of ECG
Acetylcholinesterase	ACHE	CVS, GI, pulmonary	Insufficient information	decrease in BP; decrease in HR; increase in GI motility (decrease at high doses); bronchoconstriction; increase in respiratory secretions
Cyclooxygenase 2	PTGS2	Immune, CVS	Insufficient information	Anti-inflammatory activity; anti-mitogenic effects; myocardial infarction; increase in BP; ischaemic stroke; atherothrombosis
Monoamine oxidase A	MAOA	CVS, CNS	Insufficient information	increase in BP when combined with amines such as tyramine; DDI potential; dizziness; sleep disturbances; nausea
Phosphodiesterase 3A	PDE3A	CVS	Insufficient information	increase in cardiac contractility; increase in HR; decrease in BP; thrombocytopaenia; ventricular arrhythmia
Noradrenaline transporter	SLC6A2	CNS, CVS	Insufficient information	increase in HR; increase in BP; increase in locomotor activity; constipation; abuse potential
Serotonin transporter	SLC6A4	CNS, CVS	Insufficient information	increase in GI motility; decrease in upper GI transit; decrease in plasma renin; increase in other serotonin-mediated effects; insomnia; anxiety; nausea; sexual dysfunction

Abbreviations: **HR** heart rate; **BP** blood pressure; **CO** cardiac output.

Table 2: Cardiovascular targets from Table 1 in reference [23]. Those also in Table 1 above are highlighted.

Target	Gene	Possible ADRs
Adenosine A ₁	ADORA1	Bradycardia, atrioventricular block. Renal vasoconstriction.
Adenosine A _{2A}	ADORA2A	Hypotension, coronary vasodilation. Facilitation of platelet aggregation.
Adenosine A ₃	ADORA3	Enhanced mediator release could exacerbate asthma and allergic conditions.
Adrenergic α_{1A}	ADRA1A	Hypertension and positive inotropic effect. Orthostatic hypotension.
Adrenergic α_{1B}	ADRA1B	Orthostatic hypotension.
Adrenergic α_{2A}	ADRA2A	Might inhibit insulin secretion, resulting in hyperglycemia. Hypertension exacerbates heart failure.
Adrenergic α_{2B}	ADRA2B	Hypertension, cardiac ischemia (block), vasoconstriction of arteries. Peripherally exacerbates heart failure, centrally reduces blood pressure.
Adrenergic α_{2C}	ADRA2C	Hypertension, cardiac ischemia. Increased muscular, skeletal blood flow.
Adrenergic β_1	ADRB1	Positive inotropic and chronotropic effects, ventricular fibrillation. Facilitation of bronchospasm, impairs cardiovascular performance.
Adrenergic β_2	ADRB2	Facilitates cardiac arrest, bronchodilation. Increased bronchospasm, impairs exercise stress cardiovascular performance.
Angiotensin II AT ₁	AGTR1	Increases blood pressure, cell proliferation and migration, tubular Na ⁺ resorption.
Bradykinin B ₁	BDKRB1	Enhances nociception, inflammation, vasodilation and cough.
Bradykinin B ₂	BDKRB2	Enhances nociception, inflammation, vasodilation and cough.
CGRP	CALCRL	Hypocalcaemia and hypophosphatemia.
Ca channel type L	CACNA1C	Hypotension.
Dopamine D ₁	DRD1	Treatment of Parkinson's disease; induces dyskinesia, extreme arousal, locomotor activation, vasodilatation and hypotension. Schizophrenia, neurodegeneration, coordination disorders.
Endothelin ET _a	EDNRA	Might cause vasoconstriction, positive inotropy, cell proliferation (e.g. smooth muscle and mesangial cells) and aldosterone secretion.
Endothelin ET _b	EDNRB	Causes initial vasodepression, vasoconstriction, bronchoconstriction and cell proliferation. Vasodilatation, platelet aggregation.
Ghrelin GHSR	GHSR	Energy homeostasis, GH release, effects on glucose homeostasis, cardiovascular effects.
Histamine H ₃	HRH3	Impairs memory, causes sedation, vasodilatation, bronchodilation, negative chronotropy and reduces gastrointestinal motility.
Muscarinic M ₁	CHRM1	Vagal effects, blood pressure changes, secretory functions. Decreases gastric acid secretion.
Muscarinic M ₂	CHRM2	Vagal effects, blood pressure changes. Tachycardia.
Muscarinic M ₃	CHRM3	Vagal effects, blood pressure changes, salivation. Reduces incontinence, bronchoconstriction and gastrointestinal motility. Interferes with ocular accommodation, dry mouth.
Muscarinic M ₄	CHRM4	Vagal effects, blood pressure changes. Facilitation of D1 CNS stimulation.
NE transporter	SLC6A2	Inhibitor increases adrenergic hyperactivity and facilitate α_1 adrenergic activation.
Nicotinic acetylcholine	CHRNA1	Stimulates autonomic cardiovascular, gastrointestinal functions. Palpitation, orthostatic hypotonia, nausea, sweating, muscle tremor, bronchial secretion. Effects on muscular and vegetative ganglionic functions.
NPY Y ₁	NPY1R	Antidepressant, causes vasoconstriction (venous), inhibits gut motility, gastric emptying, acid secretion, pancreatic exocrine secretions. Anxiogenic, inhibits ischemic brain injury.
K channel (hERG)	KCNH2	QT interval (electrocardiogram) prolongation.
K channel [ATP]	KCNJ11	Hypotension. Hypoglycemia.
5-HT _{2B}	HTR2B	Cardiac valvulopathy.
5-HT ₄	HTR4	Facilitates gastrointestinal transit, mechanical intestinal allodynia. Useful in treatment of irritable bowel syndrome, cardiac arrhythmias.
Na channel (site 2)	SCN5A	Antagonist causes cardiac arrhythmia.
Thromboxane A2 TP	TBXA2R	Facilitates vascular, uterine and bronchial constriction, gastrointestinal spasm, allergic inflammation and platelet aggregation. Useful in treatment of chronic productive cough, thrombosis, atherosclerosis.
Vasopressin V _{1A}	AVPR1A	Vasopressor.
Vasopressin V _{1B}	AVPR1B	Vasopressor, anxiogenic.

Also interesting is the set of anti-targets offered for screening by the contract screening company Cerep in their ADR Panel [26]. Here, the targets are annotated with the organ(s) affected, although no interaction mode (*e.g.* agonism vs. antagonism) is given. The set is described as being compiled from ADR databases, literature review and statistical association of targets with ADRs using data generated in-house [26, 52]. Although no other details are provided, the panel has been offered for some time so is presumed to have some level of acceptance within the industry. Targets described as involved in ADRs affecting the cardiovascular system are shown in Table 3; those included in either of the two sets above (*i.e.* in Tables 1 and 2 above) are highlighted.

Table 3: Cardiovascular targets taken from reference [26]. Note that in some cases the names given were unclear and gene names were assigned after inspecting the assay details in the catalogue.

Name	Gene	Name	Gene
5-HT transporter	SLC6A4	COX2	PTGS2
5-HT _{2B}	HTR2B	D ₁	DRD1
5-HT _{2C}	HTR2C	D _{4.4}	DRD4
5-HT _{4e}	HTR4	delta2 (DOP)	OPRD1
5-HT ₇	HTR7	GSK3a	GSK3A
A _{2B}	ADORA2B	H ₂	HRH2
ACE	ACE	hERG	KCNH2
acetylcholinesterase	ACHE	kappa (KOP)	OPRK1
adenylyl cyclase	ADCY5	M ₂	CHRM2
alpha _{1A}	ADRA1A	MAO-A	MAOA
alpha _{2B}	ADRA2B	MT ₃ (ML2)	MTNR1A, MTNR1B
AR	AR	Na ⁺ site 2	SCN5A
AT ₁	AGTR1	NE transporter	SLC6A2
ATPase (Na ⁺ /K ⁺)	ATP1A1-4 & ATP1B1-4	PDE3A	PDE3A
beta ₁	ADRB1	tyrosine hydroxylase	TH
Ca ²⁺ L (diltiazem site)	CACNA1C	UT	UTS2R

Again, the new targets are often members of families already seen, *e.g.* the serotonin, adenosine and dopamine receptors. Others are novel, however. The literature shows these targets mostly do have roles in the cardiovascular system, although the potential for toxicity is not always obvious.

- Although the 5-HT_{2B} isoform is a cardiovascular antitarget, there is no evidence that 5-HT_{2C} presents such a risk [53];
- The 5-HT₇ receptor plays a role in smooth muscle relaxation, and so ligands might be expected to have cardiovascular effects [54];
- The Adenosin A_{2a} receptor affects cardiac contractility, so a role in cardiotoxicity is plausible [55];
- Angiotensin Converting Enzyme (ACE) is involved in regulating vascular tone; however, ACE inhibitors are well studied in the clinic and at worst have a slight risk of inducing hypotension [56];
- Adenylate Cyclase plays an important role in regulating cardiac iontropy and lusitropy, so inhibitors could plausibly have deleterious effects on the heart [57];
- Androgens are known to mediate cardiomyocyte hypertrophy, so cardiotoxic effects from androgen receptor (AR) agonists in particular are plausible [58];
- Na⁺/K⁺-ATPase is involved in maintaining cardiomyocyte membrane potential; see the discussion in the 'Ion Channels & Pumps' section below.
- Clozapine is somewhat selective for the dopamine D4 subtype over other subtypes [59], and is associated with cardiotoxicity in the clinic [60]. Although the mechanisms of cardiotoxicity aren't clear, D4 is known to be expressed in the heart [61];
- GSK3α is involved in several cardiac signal transduction pathways; see discussion in the 'Kinases' section below;
- Melatonin receptor agonists have been shown to have cardiovascular effects, although the evidence for actual toxicity is weak. Note that MT₁ & MT₂ are the important isoforms in humans, *not* MT₃ as suggested by the Cerep list [62];

- Tyrosine hydroxylase can affect norepinephrine levels in cardiac tissue, although the evidence for toxic effects of TH inhibition is lacking [63];
- Urotensin II known to modulate cardiovascular function is a number of ways, so toxic CVS effects from urotensin II receptor (UT) ligands is plausible [64].

There do thus seem to be plausible mechanisms by which several of the novel anti-targets might induce CVS toxicity, in particular Na^+/K^+ -ATPase blockers, GSK3 α inhibitors and AR and UT agonists (although more confirmatory *in vivo* and/or clinical data would be valuable). The other targets seem to be somewhat more speculative, however, and further research on their utility is required.

A pragmatic way of treating these, and any other more speculative targets that may be encountered, would be to include them on a 'long-list', but to prioritize them below those that are judged to have a more solid involvement in toxicity in data gathering or modelling efforts.

Ion Channels & Pumps

As mentioned above, the hERG potassium channel is the most studied of the cardiac ion channels, and the one most commonly associated with arrhythmia [45]. However, various other members of the ion channel superfamily [65] are also implicated in cardiac ADRs [66]; for example, the L-type Calcium channel ($\text{Ca}_v1.2$) and the Sodium channel ($\text{Na}_v1.5$) are included alongside hERG ($\text{K}_v11.1$) on all three lists above, showing how important they are believed to be. Indeed, simulation of the effects of blockage of these three channels together has been shown to improve prediction of the risk of Torsades des Points arrhythmias over what is possible considering hERG alone [67].

The ATP-sensitive K^+ channel ($\text{K}_{ir}6.2$) and $\text{K}_v\text{LQT1}$ channel ($\text{K}_v7.1$) are also included on one pharmaceutical company list each, showing these targets are recognized there as having some degree of safety liability. Given the above, it makes sense to include the full complement of cardiac ion channels, with higher priority being given to those discussed above. The α -subunits of these channels are shown in Table 4.

Table 4: Taken from Table 1 in reference [66]. Higher priority targets are highlighted.

Current	Description	AP Phase	Activation Mechanism	Clone	Gene(s)
INa	Sodium current	Phase 0	Voltage, depolarization	Nav1.5	SCN5A
ICa,L	Calcium current, L-type	Phase 2	Voltage, depolarization	Cav1.2	CACNA1C
ICa,T	Calcium current, T-type	Phase 2	Voltage, depolarization	Cav3.1/3.2	CACNA1G,CACNA1H
Ito,f	Transient outward current, fast	Phase 1	Voltage, depolarization	KV 4.2/4.3	KCND2,KCND3
Ito,s	Transient outward current, slow	Phase 1	Voltage, depolarization	KV 1.4/1.7/3.4	KCNA4,KCNA7,KCNC4
IKur	Delayed rectifier, ultrarapid	Phase 1	Voltage, depolarization	KV 1.5/3.1	KCNA5,KCNC1
IKr	Delayed rectifier, fast	Phase 3	Voltage, depolarization	HERG ($\text{K}_v11.1$)	KCNH2
IKs	Delayed rectifier, slow	Phase 3	Voltage, depolarization	$\text{K}_v\text{LQT1}$ ($\text{K}_v7.1$)	KCNQ1
IK1	Inward rectifier	Phase3&4	Voltage, depolarization	Kir 2.1/2.2	KCNJ2,KCNJ12
IKATP	ADP activated K^+ current	Phase1&2	[ADP]/[ATP] increase	Kir 6.2 (SURA)	KCNJ11
IKAch	Muscarinic-gated K^+ current	Phase 4	Acetylcholine	Kir 3.1/3.4	KCNJ3/5
IKP	Background current	All Phases	Metabolism, stretch	TWK-1/2,TASK-1,TRAAK	KCNK1,KCNK6,KCNK3,KCNK4
If	Pacemaker current	Phase 4	Voltage, hyperpolarization	HCN2/4	HCN2,HCN4

Abbreviations: AP = Action Potential.

The Na^+/K^+ -ATPase pump is included on the Cerep list (see Table 3), which implies it is associated with some degree of safety liability. Given this fact, it is also possible that cardiac calcium pumps might also be worth including, as they too have a role in cardiac contractility [68]. However, it is also possible that intracellular targets such as the sarcoplasmic reticulum calcium pumps might not be exposed to compounds to the same extent as channels or pumps in the plasma membrane, depending on the compound's permeability.

That different target or anti-targets might experience different compound exposures depending on differences in location is an important issue, which might need to be borne in mind when interpreting the results of *in vitro* experiments in particular [69]. For example, a compound might bind tightly to a particular target, but have little effect if the free concentration at the target is low.

Kinases

The cardiotoxicity of certain protein kinase inhibitors has emerged as an issue in the field of oncology [70, 71]. A particular problem is that some of the pathways that regulate cancer cell survival are also involved in cardiomyocyte homeostasis and survival. Thus, the toxicity of anti-cancer compounds targeting these pathways is inextricably linked with the desired therapeutic mechanism, *i.e.* it is an ‘on-target’ effect. In the context of the treatment of a life-threatening cancer, this might be a tolerable risk, and several such compounds are indeed used successfully in the clinic. However, if the disease is less serious, the dosing is chronic and/or there is pre-existing cardiovascular disease then the acceptable risk will be lower. In this case, off-target perturbation of these pathways by insufficiently-selective kinase inhibitors could become a problem [72].

A recent review lists over thirty kinases believed to be of importance in the heart and vasculature, based on the results of various mouse models [73]; these are shown in Table 5. Deciding exactly which of these are important anti-targets is not straightforward. A provisional (and somewhat subjective) short list might be:

- VEGFR and PDGFR β : important in the heart’s response to stress;
- PI3K/AKT pathway: regulates cardiomyocyte survival, with AKT particularly important [74];
- CaMK II: regulates calcium homeostasis;
- AMPK: regulates cellular energy metabolism;
- GSK3 α/β : involved in regulating cardiomyocyte growth and stress response.

In light of the importance of mitochondria to this project (see section below), it is interesting to note that many kinase signalling pathways target mitochondria, and inhibition of these pathways may thus have effects on energy metabolism and cell survival [75], in the heart and elsewhere.

It must be remembered that this is an emerging area and more information is needed before a definitive list of kinase cardiovascular anti-targets can be created. In addition, it is possible that inhibition of multiple kinases might be needed to cause a toxic insult to occur, just as in some cases it is required for a therapeutic effect [76]. In this case, the kinase inhibition *profile* of a compound might be more important for toxicity than its activity at any particular kinase [73, 77].

Other target classes

There is a complement of transporters expressed in the heart [78], and there is some evidence that interaction with these transporters can be associated with cardiac toxicities [46]. There is relatively little information about this area, however, and it will not be pursued further at present.

A comprehensive list of 233 proteins linked to cardiovascular diseases in the literature has been published [79]. Many of the antitargets discussed above are included, especially amongst the GPCRs and ion channels. What is particularly interesting, however, is that many enzymes of various classes are also included. While not all of these will be relevant for cardiotoxicity, this list could provide an excellent starting point for investigating further cardiovascular antitargets.

Finally, mitochondria are known to be important in cardiotoxicity [31, 32], which opens up a range of possible anti-targets. This is discussed in a separate section, ‘Mitochondria’, below.

Table 5: Taken from Table 2 in Reference [73]. Some kinases believed to be particularly important as cardiovascular anti-targets are highlighted.

Kinase(s)	Gene(s)	Role of kinase in heart/vasculature
RAF1/BRAF	BRAF	Anti-apoptotic; preserves LV function under stress. KO: LV dysfunction and HF in the absence of additional stress; DNTG: reduced hypertrophy but LV dysfunction due to cell death

PI3K (p110α)	PIK3CA	Physiological heart growth; cardiomyocyte survival
PI3K (p110γ)	PIK3CG	Regulates contractility and pathological hypertrophy
PDK1	PDK1	Cardiomyocyte survival and β-adrenergic responsiveness
AKT1, 2 or 3	AKT1/2/3	Regulators of cardiomyocyte survival, growth and metabolism
mTOR	MTOR	mTORC1 regulates protein synthesis, inhibition leads to energy preservation under stress; mTORC2 regulates AKT activation
AMPK	PRKAA1/2, PRKAB1/2, PRKAG1/2/3	Sensor of energy stress; inhibits mTORC1, preserving energy stores. KO of AMPKα2 increased hypertrophy and LV dysfunction after TAC
GSK3α/β	GSK3A/B	Together with AMPK, inhibits mTORC1; deletion of GSKβ protective in post-MI remodelling; deletion of GSK3α leads to HF in setting of stress
CDKs	CDK2/4	CDK2 inhibition reduces ischaemia–reperfusion injury, mediated via effects on retinoblastoma protein
Aurora kinases	AURKA/B/C	M phase regulators
PLKs	PLK1	PLK1 involved in activation of CDC2, chromosome segregation, centrosome maturation, bipolar spindle formation and cytokinesis
PDGFRs	PDGFRA/B	β isoform is crucial in angiogenesis and heart's response to PO
VEGFRs	FLT1, KDR, FLT4	Crucial in angiogenesis and the heart's response to PO; antihypertensive effects
EGFR (ERBB1)	EGFR	Helps to maintain LV function in setting of chronic catecholamine stimulation; mediates pro-survival signalling
ERBB2	ERBB2	Cardiomyocyte survival and homeostasis; maintenance of LV function
KIT	KIT	Promotes CSC and immature cardiomyocyte differentiation; promotes homing to sites of MI, promoting repair.
ABL/ARG	ABL1	Maintains ER homeostasis. LV dysfunction is seen in rodents treated with imatinib
JAK2	JAK2	JAK2 and STAT3 protective in many pathological settings
FAK	PTK2	Antihypertrophic and antifibrotic in heart
DMPK	DMPK	Myotonic dystrophy type 1 is caused by excess repeats of the 3' UTR region of DMPK
LTK	LTK	Activation of LTK results in cardiac hypertrophy and cardiomyocyte degeneration
ROCK	ROCK1/2	Pro-fibrotic and pro-apoptotic in the setting of PO
LKB1	STK11	Activates AMPK which is pro-angiogenic in heart
ERK1/2	MAPK3/1	Generally promotes survival and may modulate physiological (but not pathological) hypertrophy
PKCα	PRKCA	Adverse effects on heart in setting of PO
PKG	PRKG1	One of the four nodal kinases in HF; activated by PDE5 inhibitors; inhibits apoptosis, hypertrophy and β-adrenergic responses
PIM Kinase	PIM1	Pro-survival; activated by AKT; regulated at level of gene expression
CAMKII	CAMK2A	Nodal kinase in HF; pro-hypertrophic; promotes decompensation in setting of PO Mechanism of cardiotoxicity involves regulation of CAMKII gene expression and Ca ²⁺ handling
GRK2, GRK5	ADRBK1, GRK5	Downregulates β-adrenergic signalling through recruitment of β-arrestin
ASK1	MAP3K5	Promotes pathological hypertrophy and remodelling; pro-apoptotic

Hepatotoxicity

The state of knowledge on hepatotoxicity is rather different to that of cardiotoxicity. While the understanding of basic mechanisms is growing [2, 3], there seem to be fewer unambiguously defined molecular anti-targets in the sense that has been used here. For example, in the consensus panel of 44 core safety targets mentioned above [24], 30 are annotated as having the cardiovascular system as an affected organ (see Table 1), but there are no annotations for the liver. Similarly, in an effort by the FDA to match modes of action (MOA) to adverse effects, the number of cardiac mechanisms [21] identified far outweighed the hepatobiliary mechanisms [20].

This difference presumably reflects the unique function of the liver: its role in the clearance of xenobiotics means it is exposed to high levels of reactive metabolites [80, 81], and these are one of the key drivers of drug-induced liver damage. These reactive species may exert their effects through various mechanisms, such as depletion of glutathione and covalent binding to proteins, lipids and nucleic acids [82]. This covalent binding is generally considered to be non-specific as compared to

typical non-covalent interactions, although there are attempts to document and interpret those proteins that are affected [83, 84].

Covalent modification of proteins can trigger apoptosis *via* the intrinsic pathway, or possibly necrosis in severe cases [2]. Covalent modification of proteins can also lead to haptensisation and thus to activation of the immune system [85, 86], possibly leading to liver damage *via* activation of the extrinsic apoptotic pathway. This immunogenic DILI is particularly hard to predict, as it may only manifest in susceptible individuals and is often only apparent post-marketing.

Direct interaction of parent drug or metabolites with mitochondria can also lead to cell death [29]. As mentioned above, mitochondria are important for hepatotoxicity [29, 30] as well as cardiotoxicity, and will be discussed further below.

Transporters

One class of molecular targets in the liver that are particularly important in drug discovery is the transporters [87]. One transporter known to be associated with direct hepatotoxicity is the Bile Salt Export Pump (BSEP), located in the canalicular membrane of hepatocytes. Inhibition of this transporter can result in a build-up of bile acids (BA) in hepatocytes and hence to cholestasis [88, 89]. However, other transporters in the liver are also involved in BA homeostasis and enterohepatic recirculation; four that are believed to be particularly important [90, 91] are shown in Table 6.

Table 6: Some liver transporters important in bile acid homeostasis, from references [90, 91].

Name	Gene	Location	Function
NTCP	SLC10A1	hepatocyte basolateral membrane	extracts BAs from portal blood
BSEP	ABCB11	hepatocyte canalicular membrane	secretes BAs into biliary tract
ASBT	SLC10A2	cholangiocyte apical membrane	extracts BAs from biliary tract
OST α /OST β	SLC51A/B	cholangiocyte basolateral membrane	secretes Bas back into blood

Mutations in MPR2 and MDR3 (as well as BSEP) are implicated in some hereditary cholestatic diseases [92], which would imply they too should be considered as anti-targets.

Beyond this core set, a variety of other transporters are known to have roles in bile handling in the liver [93]; some of these are listed in Table 7, with location and functional annotation. Although the focus here is on the liver, it should also be borne in mind that BA transport also occurs in tissues other than the liver, most notably the ileum and kidney [90]. This might need to be taken into account when, for example, interpreting *in vivo* data or building PK/PD models.

In addition to these transporters, BA homeostasis also relies on the hepatic Na⁺/K⁺-ATPase for maintaining the Na⁺ gradients on which some transporters, such as NTCP, rely; the hepatic CFTR channel is also required, albeit indirectly, for the functioning of some OATP transporters [93].

Table 7: Taken from Table 1 in reference [93]. Some important examples are highlighted, but note that the OST α /OST β heterodimer is not on this list as its role was discovered relatively recently.

Name	Gene	Location(s)	Main function(s)
NTCP	SLC10A1	BH	Main carrier for Na ⁺ -dependent uptake of conjugated bile salt from portal blood.
OATPs	SLCO1B1/1B3/2B1	BH	Na ⁺ -independent uptake of unconjugated bile salts and other organic anions. Polyspecific transporters with overlapping substrate affinity that are able to uptake endo- and xeno-biotics.
OCT	SLC22A1	BH	Hepatic uptake of hydrophilic organic cations. Relevant for drug transport.
OATs	SLC22A7/9	BH	Na ⁺ -independent transport of para-aminohippurate, salicylate, acetylsalicylate and methotrexate.
MRP3	ABCC3	BH, BC	Basolateral efflux of biliary constituents including non-sulfated and sulfated bile salts. Preferentially transports glucuronides but not glutathione, S-conjugates or free glutathione. Might play a role in the removal of bile acids from the liver in cholestasis.
MRP4	ABCC4	BH, BC	Mediates glutathione efflux from hepatocytes into blood by co-transport with monoanionic bile salts. Might also function as an overflow pathway during

			cholestasis. In bile duct cells, might facilitate the return of bile salts from the obstructed bile ducts to the systemic circulation.
MDR1	ABCB1	CH	ATP-dependent excretion of bulky organic cations into bile.
MDR3	ABCB4	CH	Translocation of phosphatidylcholine from inner to outer leaflet of the membrane bilayer. Crucial for biliary phospholipid secretion.
MRP2	ABCC2	CH	Canalicular conjugate export pump previously known as cMOAT. Transports bilirubin diglucuronide, sulfates, glutathione conjugates and various organic anions into bile in an ATP-dependent manner.
BSEP	ABCB11	CH	Mediates ATP-dependent bile salt transport into bile.
ABCG5	ABCG8	CH	'Half ABC transporters' that function as heterodimers to transport sterols into bile. They might also partially mediate biliary cholesterol secretion.
BCRP	ABCG2	CH	'Half ABC transporter' that mediates cellular extrusion of sulfated conjugates.
AE2	SLC4A2	CH, AC	Facilitates bicarbonate secretion into bile and contributes to bile-salt-independent bile flow.
ASBT	SLC10A2	AC	Identical to the ileal bile salt transporter. Functions as an uptake mechanism for bile salts, removing them from bile.
FIC1	ATP8B1	CH, AC	Member of the Type IV P-type ATPase family, which functions as an ATP-dependent aminophospholipid translocase. However, FIC1 function is not yet clearly defined. It is mutated in two different disorders: PFIC1 and BRIC.

Abbreviations: BH = Basolateral membrane of hepatocytes; CH = Canalicular membrane of hepatocytes; BC = Basolateral membrane of cholangiocytes; AC = apical membrane of cholangiocytes.

It would seem plausible that inhibition of one or more of these transporters could cause toxic effects due to the disruption of bile homeostasis. However, the extent to which toxicities beyond those associated with BSEP occur is not yet clear.

As well as direct toxic effects, transporters are frequently involved in drug-drug interactions. Their importance in this context is such that a consortium has been formed by various pharmaceutical companies (the ITC) in order to provide guidance on which transporters are of concern and on the methods used to study them [94, 95]. Transporters of interest to the ITC are shown in Table 8. The full list is included, to emphasize that all may be important when considering DDIs. Those expressed in the liver are highlighted, as these are presumably more likely to be important for hepatotoxicity (by whatever mechanism).

The overlap of transporters considered important for DDIs with those involved in bile acid transport is evident. It is thus conceivable that inhibitors could exert toxic effects both through disruption of BA homeostasis and through DDIs.

Table 8: Hepatic transporters from the ITC reviews [94, 95]. Those considered to be of particular relevance to drug discovery are bolded.

Name	Gene
OATP1B1	SLCO1B1
OATP1B3	SLCO1B3
OATP1A2	SLCO1A2
OATP2B1	SLCO2B1
OCT1	SLC22A1
MATE1	SLC47A1
MATE2-K	SLC47A2
MDR1	ABCB1
BCRP	ABCG2
BSEP	ABCB11
MRP2	ABCC2
MRP3	ABCC3
MRP4	ABCC4
MDR3	ABCB4
OAT2	SLC22A7
OAT7	SLC22A9

NTCP	SLC10A1
OST α/β	SLC51A/B
MRP6	ABCC6
ENT1/2	SLC29A1/2

It is interesting in light of the discussion of mitochondria below that there is a complement of transporters largely specific to the mitochondria [96, 97]. The impermeability of the inner mitochondrial membrane means transporters are crucial for the import and export of metabolites, and it is possible interference with these processes could affect mitochondrial function, in the liver and other organs.

Other transporters beyond those mentioned above are known to be expressed in the liver and elsewhere [78]; however, evidence any involvement in toxicity or DDIs is generally lacking and they will not be considered further at present.

Note that a compound may block a transporter or be a substrate for it. It is possible that both modes could lead to toxic effects, albeit by different mechanisms: the first perhaps by altering the disposition of another drug or of an endogenous molecule, the second by altering the disposition of the transported drug. This distinction would need to be kept in mind when gathering, interpreting and modelling drug/transporter data.

Nuclear Receptors

Nuclear receptors (NR) control expression of ADME proteins (XMEs, transporters *etc.*) in response to xenobiotic insult. Interaction of drugs with NRs can therefore have deleterious consequences, most notably DDIs caused by induction of XMEs [98]. This induction can also cause direct liver toxicity, perhaps by increasing the concentration of reactive metabolites and/or reactive oxygen species [3].

However, NRs are also involved with many homeostatic processes in the liver, including bile acid metabolism/bile secretion [99, 100] and lipid metabolism [101]. Because of these regulatory roles, there is interest in these receptors as therapeutic targets for cholestasis and steatohepatitis. However, as with other therapeutic targets, it is conceivable that inappropriate modulation of these receptors (*e.g.* in different disease states) could lead to toxic outcomes and that off-target interaction with NRs is best avoided. Thus, pharmaceutical companies typically screen against some subset of NRs [102], and CROs offer various NR assays [25, 26]. Nuclear receptors with roles in the liver are listed in Table 9, with those judged most likely to be involved in hepatotoxicity highlighted [99-101]. Note that the transcription factors AhR and Nrf1 are not NRs, but are often discussed alongside them because of their closely related roles.

Table 9: Nuclear Receptors with roles in the liver [99]; those believed to be most important for hepatotoxicity are highlighted [100, 101].

Name	Systematic Name	Gene
FXR	NR1H4	NR1H4
SHP	NR0B2	NR0B2
PXR	NR1I2	NR1I2
CAR	NR1I3	NR1I3
VDR	NR1I1	VDR
HNF4 α	NR2A1	HNF4A
LRH1	NR5A2	NR5A2
PPAR α	NR1C1	PPARA
PPAR γ	NR1C3	PPARG
LXR α	NR1H3	NR1H3
LXR β	NR1H2	NR1H2
GR	NR3C1	NR3C1
RAR α	NR1B1	RARA

AhR	n/a	AHR
Nrf2	n/a	NFE2L2

Note that, as with other types of receptors, ligands for NRs can be agonists, antagonists and possibly inverse agonists. Proper annotation of compound-receptor activity data with the mode of interaction is thus crucial for proper interpretation and modeling.

Other targets

The Cerep ADR Panel [26] contains targets flagged as relevant to liver toxicity as well as heart toxicity, although the lists overlap heavily. The liver anti-targets are shown in Table 10.

Table 10: Liver targets taken from reference [26]. Those that are also cardiovascular anti-targets are highlighted.

Name	Gene	Name	Gene
ACE	ACE	ERK2 (P42mapk)	MAPK1
alpha2A	ADRA2A	ETB	EDNRB
AR	AR	GR	NR3C1
ATPase (Na ⁺ /K ⁺)	ATP1A1-4, ATP1B1-4	GSK3a	GSK3A
beta2	ADRB2	H2	HRH2
carbonic anhydrase II	CA2	MAO-A	MAOA
constitutive NOS (endothelial)	NOS3	motilin	MLNR
COX2	PTGS2	Na ⁺ site 2	SCN5A
D4.4	DRD4	NMDA	GRIN1
ERalpha	ESR1		

As with the cardiovascular case, some of these anti-targets or their families have already been identified as being of interest, while others are novel. Some examples are discussed below:

- Liver injury from ACE inhibitors has been reported but it is rare and there does not seem to be evidence that it is an on-target effect [103];
- Although there is some evidence that α_2A receptor agonists might potentiate hepatotoxicity of co-administered xenobiotics [104], there seems to be little or no evidence on-target hepatotoxicity associated with this receptor;
- Androgen signalling *via* the androgen receptor (AR) has been shown to suppress the development of steatosis [105]; it is thus plausible that inhibition of this process by xenobiotics might be deleterious, at least in some disease states;
- The Na⁺/K⁺-ATPase, as noted above, is involved in bile-acid homeostasis [93]. It is thus plausible that interference with the activity of this pump could lead to liver damage;
- Although there is some evidence that β_2 receptor agonists might potentiate hepatotoxicity of co-administered xenobiotics [104], there seems to be little evidence of direct DILI associated with either agonists [106] or antagonists [107]. As these are very widely used classes of drugs, any on-target hepatotoxicity should be readily apparent; that it is not suggests that this target is unlikely to be a genuine DILI liability;
- Some carbonic anhydrase (CA) inhibitors, such as acetazolamide, have been associated with idiosyncratic DILI, albeit rarely [108]. However, this appears to be a class effect associated with the thiadiazolone moiety as opposed to on-target toxicity. In addition, CA inhibitors are contraindicated for patients with severe liver disease as the diuretic effect can lead to hypokalaemia and hence induce coma [109]. While this could be important clinically, it does not mean CA inhibitors are likely to be hepatotoxic in non-vulnerable patient populations;
- Nitric oxide signaling is important in the liver [110], so it is conceivable that inhibition of eNOS could be detrimental in some circumstances;
- NSAIDs are very widely used drugs and, while cases of idiosyncratic hepatotoxicity have been reported [111], there does not seem to be much evidence of on-target hepatotoxicity associated with COX2;

- There does not appear to be clinical evidence of on-target hepatotoxicity for dopamine receptor ligands in general. Compounds selective for the D4 isoform are uncommon, so data here is particularly scarce. However, apomorphine is a fairly potent D4 agonist (also active at other isoforms) that is, at most, rarely associated with DILI [112];
- Estrogen receptor α (ER α) is expressed in the liver, although, by contrast with the AR, there does not seem to be any obvious connection with liver disease [113];
- The kinases ERK2 and GSK3 α and plausible hepatic anti-targets, as they are involved in cell growth and survival pathways. Hepatocytes exist in a high-stress environment, so might well be vulnerable to disruption of these pathways. It is notable in this context that these appear on the list of cardiac kinases of concern [73];
- There is some evidence of a risk of hepatotoxicity in patients receiving Endothelin receptor antagonists [114]. However, there seems to be little or no evidence that this is due to on-target activity at ETB: a more plausible suggestion is that the sulphonamide moiety in Bosentan, the most common ET inhibitor in use, might be responsible for the observed DILI;
- The glucocorticoid receptor (GR) is a known anti-target, as noted above;
- Histamine H2 receptor antagonists are extremely widely used and well tolerated [115]: there seems to be no reason to suspect on-target hepatotoxicity in this case. Relatively little information on agonists is available;
- Monoamine Oxidase (MAO) inhibitors have historically been associated with DILI; however, this is likely to have been due to reactive metabolites related to the presence of a hydrazine moiety in many earlier examples, and not to on-target hepatotoxicity [116];
- There seems to be little evidence linking motilin receptor ligands to hepatotoxicity. The macrolide antibiotic erythromycin is associated with a low rate DILI, which is believed to occur *via* an allergic mechanism; however, the frequency of use means such cases are not uncommon [117]. A hemiketal metabolite of drug is a motilin receptor agonist, responsible for the well-known GI side-effects [118]. However, there is no evidence linking this activity to the hepatotoxicity;
- The epithelial sodium channel ENaC works with the Na⁺/K⁺-ATPase pump in maintaining sodium ion homeostasis [119], which as has been noted above, is important in bile acid homeostasis. However, this assay is based on the voltage-gated sodium channel. While the multi-target antiarrhythmic drug Dronedaron has some activity at this channel, and has a mild association with clinical DILI [120], there is no reason to believe this is due to the sodium channel activity;
- There seems to be no evidence that activity at the NMDA receptor is responsible for DILI. The NMDA receptor antagonist Memantine is only rarely associated with liver damage in clinical use, and there is no reason to believe this is an on-target effect [121].

Thus, it appears that some of these targets do appear to be legitimate hepatotoxicity anti-targets, and for others there is at least a plausible mechanism by which they might be involved in DILI. Some caution should be exercised in the latter cases, however. For example, the involvement of CA inhibitors in hepatotoxicity is genuine but indirect and these compounds are unlikely to induce DILI in those without serious preexisting liver disease. It might thus be somewhat misleading to characterize such a target as a hepatotoxicity anti-target.

Furthermore, for several targets here the connection with hepatotoxicity is, at best, very weak; this seems to be particularly the case with the GPCRs in the list. Several of these are targets of very widely used drugs, and any on-target hepatotoxicity should be very evident. However, a sufficiently large patient population means there may be many cases of idiosyncratic toxicity observed, even though the actual rate is very low. In addition, idiosyncratic toxicities are generally not on-target effects of a parent drug, but mediated by metabolites through a variety of mechanisms [122]. Thus, while there may be many cases of DILI recorded against a drug with a particular therapeutic target, that target may not in reality be associated with hepatotoxicity in any meaningful way.

In conclusion, this list of hepatotoxicity anti-targets is interesting due to the mechanistic diversity it suggests. However, in many cases, further literature research would be required to find evidence

corroborating their involvement with hepatotoxicity and/or providing mechanistic rationales for their inclusion.

Xenobiotic Metabolising Enzymes

As noted in the introduction, XMEs are crucial in mediating various toxicities through the production of active metabolites [80, 81, 123] and possibly *via* drug-drug interactions [36]. Thus, understanding their interactions remains vital to any effort at predictive toxicology. These interactions can be diverse: compounds can be substrates (possibly generating reactive metabolites) or inhibitors of one or more enzymes; they can also act as inducers [3] *via* nuclear receptors (see above).

The PharmaADME group, with pharmaceutical industry participation, designed a ‘core list’ of 32 ADME genes designed “to identify predictors of pharmacokinetic variability that could impact drug safety and efficacy in the current drug development process” [35]. This core list includes Phase I and II XMEs and transporters; these are shown in Table 11. Note that this list is not restricted to hepatic species only, for reasons already discussed.

The group also provides an ‘extended list’ of 267 genes, intended to give a complete set of genes associated with drug metabolism [35]. As well as further XMEs and transporter isoforms, the extended list includes ‘modifiers’: these are nuclear receptors responsible for induction, ancillary enzymes such as cytochrome P450 oxidoreductase and other species required for the proper functioning of the ADME machinery.

Although these lists were designed around pharmacogenomics experiments [124], they together serve as a definitive list of ADME-related genes. For example, the core set is largely, and the extended set entirely, a superset of the ADME targets offered by screening companies [125, 126], the ITC transporters of interest [94, 95] and targets identified by regulators as involved in DDIs [36].

Table 11: Taken from reference [36].

Gene Symbol	Full Gene Name	Class
CYP1A1	cytochrome P450, family 1, subfamily A, polypeptide 1	Phase I
CYP1A2	cytochrome P450, family 1, subfamily A, polypeptide 2	Phase I
CYP2A6	cytochrome P450, family 2, subfamily A, polypeptide 6	Phase I
CYP2B6	cytochrome P450, family 2, subfamily B, polypeptide 6	Phase I
CYP2C8	cytochrome P450, family 2, subfamily C, polypeptide 8	Phase I
CYP2C9	cytochrome P450, family 2, subfamily C, polypeptide 9	Phase I
CYP2C19	cytochrome P450, family 2, subfamily C, polypeptide 19	Phase I
CYP2D6	cytochrome P450, family 2, subfamily D, polypeptide 6	Phase I
CYP2E1	cytochrome P450, family 2, subfamily E, polypeptide 1	Phase I
CYP3A4	cytochrome P450, family 3, subfamily A, polypeptide 4	Phase I
CYP3A5	cytochrome P450, family 3, subfamily A, polypeptide 5	Phase I
DPYD	dihydropyrimidine dehydrogenase	Phase I
GSTM1	glutathione S-transferase M1	Phase II
GSTP1	glutathione S-transferase pi	Phase II
GSTT1	glutathione S-transferase theta 1	Phase II
NAT1	N-acetyltransferase 1 (arylamine N-acetyltransferase)	Phase II
NAT2	N-acetyltransferase 2 (arylamine N-acetyltransferase)	Phase II
SULT1A1	sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1	Phase II
TPMT	thiopurine S-methyltransferase,	Phase II
UGT1A1	UDP glucuronosyltransferase 1 family, polypeptide A1	Phase II
UGT2B15	UDP glucuronosyltransferase 2 family, polypeptide B15	Phase II
UGT2B17	UDP glucuronosyltransferase 2 family, polypeptide B17	Phase II
UGT2B7	UDP glucuronosyltransferase 2 family, polypeptide B7	Phase II
ABCB1	ATP-binding cassette, sub-family B (MDR/TAP), member 1	Transporter
ABCC2	ATP-binding cassette, sub-family C (CFTR/MRP), member 2	Transporter

ABCG2	ATP-binding cassette, sub-family G (WHITE), member 2	Transporter
SLC15A2	solute carrier family 15 (H ⁺ /peptide transporter), member 2	Transporter
SLC22A1	solute carrier family 22 (organic cation transporter), member 1	Transporter
SLC22A2	solute carrier family 22 (organic cation transporter), member 2	Transporter
SLC22A6	solute carrier family 22 (organic anion transporter), member 6	Transporter
SLCO1B1	solute carrier organic anion transporter family, member 1B1	Transporter
SLCO1B3	solute carrier organic anion transporter family, member 1B3	Transporter

An application using this list is the ADME Safari, which integrates tissue expression, orthologues (valuable for cross-species extrapolation) and bioassay data for these proteins and provides convenient access *via* a web portal [127].

Mitochondria

Mitochondria are frequently involved in toxic responses to drugs [29-32]. This is likely to be for two main reasons. First, they play a key role in apoptosis, where they may be effectors of toxic responses triggered by initiating events in which they were not directly involved. Second, they may be targets of toxicants themselves, for example *via* disruption of the citric acid cycle, fatty-acid oxidation or the electron transport chain.

The relevance to HeCaToS is great as mitochondria are particularly important to both the liver [29, 30] and heart [31, 32]; the former because of role in energy homeostasis and the latter because of its energy requirements. In addition, the liver's role in the clearance of xenobiotics means exposure to both parent compounds and active metabolites is likely to be higher than in other tissues.

The involvement in energy homeostasis, lipogenesis *etc.* means mitochondria are of great interest to those studying metabolic diseases. This focus, alongside their roles in apoptosis and toxicity, has led to the creation of several resources integrating various types of 'omics data for mitochondria [128-130]. These resources provide a comprehensive overview of which proteins are found in mitochondria, and which might therefore be the molecular targets of mitochondrial toxins. However, as is often the case with such exhaustive lists of proteins, not all have yet been well characterized and the level of annotation varies widely. Other useful sources of information here are pathway databases such as Reactome [33] and ConsensusPathDB [131], which place individual targets in the biochemical context in which they operate.

The Reactome database was chosen as the basis for an initial investigation of the possibilities for modeling drug-induced mitochondrial dysfunction. This resource contains a particularly rich description of proteins and complexes and the reactions and pathways they are involved in. This can appear complicated when compared with the simple lists of proteins: for example, proteins can appear individually and as part of complexes, complexes might be included in both reduced and oxidized forms and in multiple locations and pathways *etc.* However, this complexity reflects the underlying biology, and provides many extra insights. In addition to enabling mitochondrial modeling, the pathway databases will be useful in further investigation of some of the other anti-targets discussed above.

As an example application, mitochondrial entities from Reactome were mapped to ChEMBL targets, using shared protein chain membership. This enabled the retrieval of ChEMBL data for compounds active against those targets, a prelude to investigating the possibilities of building QSAR models for these targets.

Difficulties

There are several related issues to be considered with regard to the information presented above. The first is that there is likely to be gaps in the coverage of anti-targets. Despite recent advances, aspects of both cardiotoxicity and hepatotoxicity remain poorly understood and this suggests that there are anti-targets and/or mechanisms of toxicity and that remain to be identified.

One way of expanding anti-target coverage might be to use pathway databases to identify other targets on the same pathways as known anti-targets. For example, if antagonism of a cell-surface receptor is known to result in toxicity in some circumstances, then it is plausible that disruption of elements of the signal transduction pathway(s) associated with that receptor might result in a similar phenotype. Similarly, if inhibition of an enzyme on a metabolic pathway results in toxicity, it is possible that inhibition of other enzymes on that pathway might give a similar effect. This sort of thinking is common when attempting to discover new therapeutic targets, and it should also be applicable to anti-target discovery.

There are problems with the approach, however. Importantly, it is likely to introduce false positives: for example, it is well known that there is redundancy built into many signalling pathways [132] and that the effects of inhibiting enzymes on a metabolic pathway might differ depending on whether they catalyse a rate-limiting step or not [133]. In an industrial target discovery setting these hypotheses could be tested by experiment, while in the current context only the literature is available. For example, quantitative models of metabolic networks might help to identify those enzymes that would cause the most disruption if inhibited [134].

Another way of expanding the range of mechanisms covered would be to consult one of the various lists of targets associated with ADRs that have been published [18-21]. In addition, toxicogenomics experiments can identify genes and associated pathways perturbed during a toxic response [135, 136], although the link to anti-targets as discussed here is not always clear.

While these data-driven approaches are attractive, they can only be a starting point in the identification of anti-targets, as confirmatory data (*in vivo* or clinical) or a convincing mechanistic hypothesis would still be required to validate an anti-target. Again, as experiment is not an option here, only the literature is available for further investigating hypotheses.

As noted previously, there are databases of toxicity-associated targets [6, 7] that are useful for annotating known anti-targets but difficult to mine systematically. A related resource which might be useful for this purpose is the Comparative Toxicogenomics Database [137], which links diseases, genes/proteins and compounds. While this began as a resource aimed at environmental toxicants, there is now more of an emphasis on drug-like compounds [138], and the ability to download data means algorithmic mining might be more practical.

Once novel candidate anti-targets are identified, the issue of their validation arises. The amount of information available on different anti-targets differs greatly, and what is actually 'sufficient' in a given context is an open question. Ideally, a comprehensive AOP would be available [9, 88], with unambiguous *in vitro*, *in vivo* or *clinical data* to support each step. However, this will not be available except in rare cases, and decisions (*e.g.* on what to model) will have to be made on incomplete information.

Another issue when considering pharmaceutical anti-targets (or targets) is that of inter-individual variation. In many cases, toxicities only appear after marketing in a small subset of patients. While this can be due to the involvement of the immune system [85, 86], it might also be due to different protein isoforms being present in different individuals [139]. While many such differences will be

silent, if they are present in the active site or in a recognition element then they could give rise to differing responses [140-142] to xenobiotics.

It should also be noted that the interactions of xenobiotics with the anti-targets discussed here are at most likely to be risk factors for toxicity. Any actual observable toxic response at a tissue, organ or organism level will most likely vary depending on various factors such as compound exposure, intra-individual genetic variation, pre-existing conditions and co-administered therapeutics.

As a practical issue, the amount of data available for various anti-targets of interest will vary widely. In some cases easily available data for anti-targets of interest may not be sufficient for QSAR model building, for example. The decision will then have to be made as to whether the target is sufficiently important to warrant possibly time-consuming or expensive data-gathering and curation activities.

References

1. Przybylak, K.R. and M.T. Cronin, *In silico models for drug-induced liver injury--current status*. Expert Opin Drug Metab Toxicol, 2012. **8**(2): p. 201-17.
2. Russmann, S., G.A. Kullak-Ublick, and I. Grattagliano, *Current concepts of mechanisms in drug-induced hepatotoxicity*. Curr Med Chem, 2009. **16**(23): p. 3041-53.
3. Anderson, N. and J. Borlak, *Mechanisms of Toxic Liver Injury*, in *Hepatotoxicity: From Genomics to in Vitro and in Vivo Models*, S.C. Sahu, Editor. 2007, Wiley-Blackwell. p. 191-286.
4. *Antitargets: Prediction and Prevention of Drug Side Effects*. Methods and Principles in Medicinal Chemistry, ed. R. Mannhold, H. Kubinyi, and G. Folkers. 2008: WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.
5. Urban, L., et al., *Screening for Safety-Relevant Off-Target Activities*, in *Polypharmacology in Drug Discovery* J.-W. Peters, Editor. 2012.
6. Zhang, J.X., et al., *DITOP: drug-induced toxicity related protein database*. Bioinformatics, 2007. **23**(13): p. 1710-2.
7. Ji, Z.L., et al., *Drug Adverse Reaction Target Database (DART) : proteins related to adverse drug reactions*. Drug Saf, 2003. **26**(10): p. 685-90.
8. Allen, T.E., et al., *Defining molecular initiating events in the adverse outcome pathway framework for risk assessment*. Chem Res Toxicol, 2014. **27**(12): p. 2100-12.
9. Vinken, M., *The adverse outcome pathway concept: a pragmatic tool in toxicology*. Toxicology, 2013. **312**: p. 158-65.
10. *In Silico Toxicology : Principles and Applications*. Issues in Toxicology. 2010: The Royal Society of Chemistry.
11. Bai, J.P. and D.R. Abernethy, *Systems pharmacology to predict drug toxicity: integration across levels of biological organization*. Annu Rev Pharmacol Toxicol, 2013. **53**: p. 451-73.
12. Jens Niklas, J., et al., *Quantitative Evaluation and Prediction of Drug Effects and Toxicological Risk Using Mechanistic Multiscale Models*. Mol Inform, 2013. **32**: p. 14-23.
13. Bhattacharya, S., et al., *Modeling drug- and chemical-induced hepatotoxicity with systems biology approaches*. Front Physiol, 2012. **3**: p. 462.
14. IUPHAR/BPS. *Guide to Pharmacology*. 2014; Available from: <http://www.guidetopharmacology.org/>.
15. Landesmann, B., et al., *Description of Prototype Modes-of-Action Related to Repeated Dose Toxicity*, 2012.
16. Muthas, D. and S. Boyer, *Exploiting Pharmacological Similarity to Identify Safety Concerns – Listen to What the Data Tells You*. Mol Inform, 2013. **32**(1): p. 37-45.
17. Peters, J.U., *Polypharmacology - foe or friend?* J Med Chem, 2013. **56**(22): p. 8955-71.
18. Kuhn, M., et al., *Systematic identification of proteins that elicit drug side effects*. Mol Syst Biol, 2013. **9**: p. 663.
19. Lounkine, E., et al., *Large-scale prediction and testing of drug activity on side-effect targets*. Nature, 2012. **486**(7403): p. 361-7.
20. Matthews, E.J., et al., *Identification of structure-activity relationships for adverse effects of pharmaceuticals in humans: Part C: use of QSAR and an expert system for the estimation of the mechanism of action of drug-induced hepatobiliary and urinary tract toxicities*. Regul Toxicol Pharmacol, 2009. **54**(1): p. 43-65.
21. Matthews, E.J. and A.A. Frid, *Prediction of drug-related cardiac adverse effects in humans--A: creation of a database of effects and identification of factors affecting their occurrence*. Regul Toxicol Pharmacol, 2010. **56**(3): p. 247-75.
22. Bowes, J., et al., *Pharmacological and Pharmaceutical Profiling: New Trends*, in *The Process of New Drug Discovery and Development*, C.G. Smith and J.T. O'Donnell, Editors. 2006, CRC Press. p. 103-134.
23. Whitebread, S., et al., *Keynote review: in vitro safety pharmacology profiling: an essential tool for successful drug development*. Drug Discov Today, 2005. **10**(21): p. 1421-33.

24. Bowes, J., et al., *Reducing safety-related drug attrition: the use of in vitro pharmacological profiling*. Nat Rev Drug Discov, 2012. **11**(12): p. 909-22.
25. Cerep. *Organ Tox Panel*. 2014; Available from: [http://www.cerep.fr/cerep/users/pages/downloads/Documents/Marketing/Pharmacology & ADME/OTP/Organ Tox Panel.pdf](http://www.cerep.fr/cerep/users/pages/downloads/Documents/Marketing/Pharmacology&ADME/OTP/OrganToxPanel.pdf).
26. Cerep. *ADR Panel*. 2104; Available from: [http://www.cerep.fr/cerep/users/pages/Downloads/Documents/Marketing/Pharmacology & ADME/OTP/ADRPanell.pdf](http://www.cerep.fr/cerep/users/pages/Downloads/Documents/Marketing/Pharmacology&ADME/OTP/ADRPanell.pdf).
27. *The Human Protein Atlas*. 2014.
28. *TISSUES: Tissue expression database*. 2014.
29. Begrich, K., et al., *Drug-induced toxicity on mitochondria and lipid metabolism: mechanistic diversity and deleterious consequences for the liver*. J Hepatol, 2011. **54**(4): p. 773-94.
30. Pessayre, D., et al., *Central role of mitochondria in drug-induced liver injury*. Drug Metab Rev, 2012. **44**(1): p. 34-87.
31. Di Lisa, F., et al., *Mitochondrial Dysfunction in Cell Injury and Cardiotoxicity*, in *Cardiotoxicity of Non-Cardiovascular Drugs*, G. Minotti, Editor. 2010, Wiley-Blackwell. p. 1-24.
32. *Drug-Induced Mitochondrial Dysfunction*. 2008: Wiley.
33. *Reactome: A curated pathway database*. 2014.
34. Wu, F., et al., *Computer modeling of mitochondrial tricarboxylic acid cycle, oxidative phosphorylation, metabolite transport, and electrophysiology*. J Biol Chem, 2007. **282**(34): p. 24525-37.
35. PharmaADME. *PharmaADME*. 2104; Available from: <http://pharmaadme.org/>.
36. Prueksaritanont, T., et al., *Drug-drug interaction studies: regulatory guidance and an industry perspective*. Aaps j, 2013. **15**(3): p. 629-45.
37. Thelen, K. and J.B. Dressman, *Cytochrome P450-mediated metabolism in the human gut wall*. J Pharm Pharmacol, 2009. **61**(5): p. 541-58.
38. Inui, K.I., S. Masuda, and H. Saito, *Cellular and molecular aspects of drug transport in the kidney*. Kidney Int, 2000. **58**(3): p. 944-58.
39. Castiglione, F., et al., *Modeling biology spanning different scales: an open challenge*. Biomed Res Int, 2014. **2014**: p. 902545.
40. Sahdeo, S., et al., *High-throughput screening of FDA-approved drugs using oxygen biosensor plates reveals secondary mitofunctional effects*. Mitochondrion, 2014. **17**: p. 116-25.
41. Persson, M., et al., *A high content screening assay to predict human drug-induced liver injury during drug discovery*. J Pharmacol Toxicol Methods, 2013. **68**(3): p. 302-13.
42. Biour, M., et al., *[Drug-induced liver injury; fourteenth updated edition of the bibliographic database of liver injuries and related drugs]*. Gastroenterol Clin Biol, 2004. **28**(8-9): p. 720-59.
43. Kovacic, P., et al., *Mechanism of mitochondrial uncouplers, inhibitors, and toxins: focus on electron transfer, free radicals, and structure-activity relationships*. Curr Med Chem, 2005. **12**(22): p. 2601-23.
44. Soares, K.M., et al., *Profiling the NIH Small Molecule Repository for compounds that generate H₂O₂ by redox cycling in reducing environments*. Assay Drug Dev Technol, 2010. **8**(2): p. 152-74.
45. Raschi, E., et al., *hERG-related drug toxicity and models for predicting hERG liability and QT prolongation*. Expert Opin Drug Metab Toxicol, 2009. **5**(9): p. 1005-21.
46. Lavery, H., et al., *How can we improve our understanding of cardiovascular safety liabilities to develop safer medicines?* Br J Pharmacol, 2011. **163**(4): p. 675-93.
47. Maxwell, C.B. and A.T. Jenkins, *Drug-induced heart failure*. Am J Health Syst Pharm, 2011. **68**(19): p. 1791-804.
48. Sobanski, P., et al., *The presence of mu-, delta-, and kappa-opioid receptors in human heart tissue*. Heart Vessels, 2014. **29**(6): p. 855-63.
49. Salazar, N.C., J. Chen, and H.A. Rockman, *Cardiac GPCRs: GPCR Signaling in Healthy and Failing Hearts*. Biochim Biophys Acta, 2007. **1768**(4): p. 1006-18.

50. Foster, S.R., et al., *G protein-coupled receptors in cardiac biology: old and new receptors*. Biophysical Reviews: p. 1-13.
51. Kaczorowski, G.J. and O. Pongs, *Editorial overview: Cardiovascular and renal: Novel therapeutic strategies and approaches for targeting unmet cardiovascular needs*. Curr Opin Pharmacol, 2014. **15**: p. v-viii.
52. Krejsa, C.M., et al., *Predicting ADME properties and side effects: the BioPrint approach*. Current opinion in drug discovery & development, 2003. **6**(4): p. 470-480.
53. Chen, G., et al., *Rational Drug Design Leading to the Identification of a Potent 5-HT(2C) Agonist Lacking 5-HT(2B) Activity*. ACS Med Chem Lett, 2011. **2**(12): p. 929-932.
54. Vanhoenacker, P., G. Haegeman, and J.E. Leysen, *5-HT7 receptors: current knowledge and future prospects*. Trends Pharmacol Sci, 2000. **21**(2): p. 70-7.
55. Chandrasekera, P.C., et al., *Differential effects of adenosine A2a and A2b receptors on cardiac contractility*. Am J Physiol Heart Circ Physiol, 2010. **299**(6): p. H2082-9.
56. Flather, M.D., et al., *Long-term ACE-inhibitor therapy in patients with heart failure or left-ventricular dysfunction: a systematic overview of data from individual patients*. ACE-Inhibitor Myocardial Infarction Collaborative Group. Lancet, 2000. **355**(9215): p. 1575-81.
57. Feldman, A.M., *Adenylyl cyclase: a new target for heart failure therapeutics*. Circulation, 2002. **105**(16): p. 1876-8.
58. Marsh, J.D., et al., *Androgen receptors mediate hypertrophy in cardiac myocytes*. Circulation, 1998. **98**(3): p. 256-61.
59. PDSP. *PDSP Ki Database*. 2015; Available from: <http://pdsp.med.unc.edu/pdsp.php>.
60. Layland, J.J., D. Liew, and D.L. Prior, *Clozapine-induced cardiotoxicity: a clinical update*. Med J Aust, 2009. **190**(4): p. 190-2.
61. Cavallotti, C., et al., *Dopamine receptor subtypes in the native human heart*. Heart Vessels, 2010. **25**(5): p. 432-7.
62. Paulis, L., F. Simko, and M. Laudon, *Cardiovascular effects of melatonin receptor agonists*. Expert Opin Investig Drugs, 2012. **21**(11): p. 1661-78.
63. Eisenhofer, G., et al., *Cardiac sympathetic nerve function in congestive heart failure*. Circulation, 1996. **93**(9): p. 1667-76.
64. Russell, F.D., *Urotensin II in cardiovascular regulation*. Vasc Health Risk Manag, 2008. **4**(4): p. 775-85.
65. Yu, F.H., et al., *Overview of molecular relationships in the voltage-gated ion channel superfamily*. Pharmacol Rev, 2005. **57**(4): p. 387-95.
66. Grant, A.O., *Cardiac ion channels*. Circ Arrhythm Electrophysiol, 2009. **2**(2): p. 185-94.
67. Mirams, G.R., et al., *Simulation of multiple ion channel block provides improved early prediction of compounds' clinical torsadogenic risk*. Cardiovasc Res, 2011. **91**(1): p. 53-61.
68. Brini, M. and E. Carafoli, *Calcium pumps in health and disease*. Physiol Rev, 2009. **89**(4): p. 1341-78.
69. Smith, D.A., L. Di, and E.H. Kerns, *The effect of plasma protein binding on in vivo efficacy: misconceptions in drug discovery*. Nat Rev Drug Discov, 2010. **9**(12): p. 929-39.
70. Mellor, H.R., et al., *Cardiotoxicity associated with targeting kinase pathways in cancer*. Toxicol Sci, 2011. **120**(1): p. 14-32.
71. Cheng, H. and T. Force, *Molecular mechanisms of cardiovascular toxicity of targeted cancer therapeutics*. Circ Res, 2010. **106**(1): p. 21-34.
72. Anastassiadis, T., et al., *Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity*. Nat Biotechnol, 2011. **29**(11): p. 1039-45.
73. Force, T. and K.L. Kolaja, *Cardiotoxicity of kinase inhibitors: the prediction and translation of preclinical models to clinical outcomes*. Nat Rev Drug Discov, 2011. **10**(2): p. 111-26.
74. Sussman, M.A., et al., *Myocardial AKT: the omnipresent nexus*. Physiol Rev, 2011. **91**(3): p. 1023-70.
75. Horbinski, C. and C.T. Chu, *Kinase signaling cascades in the mitochondrion: a matter of life or death*. Free Radic Biol Med, 2005. **38**(1): p. 2-11.

76. Chow, L.Q. and S.G. Eckhardt, *Sunitinib: from rational design to clinical efficacy*. J Clin Oncol, 2007. **25**(7): p. 884-96.
77. Olaharski, A.J., et al., *Identification of a kinase profile that predicts chromosome damage induced by small molecule kinase inhibitors*. PLoS Comput Biol, 2009. **5**(7): p. e1000446.
78. Lee, E.J., C.B. Lean, and L.M. Limenta, *Role of membrane transporters in the safety profile of drugs*. Expert Opin Drug Metab Toxicol, 2009. **5**(11): p. 1369-83.
79. Cases, M. and J. Mestres, *A chemogenomic approach to drug discovery: focus on cardiovascular diseases*. Drug Discov Today, 2009. **14**(9-10): p. 479-85.
80. Kalgutkar, A.S., et al., *A comprehensive listing of bioactivation pathways of organic functional groups*. Curr Drug Metab, 2005. **6**(3): p. 161-225.
81. Macherey, A.-C. and P.M. Dansette, *Biotransformations Leading to Toxic Metabolites: Chemical Aspects*, in *The Practice of Medicinal Chemistry*, C.G. Wermuth, Editor. 2008, Academic Press. p. 674-696.
82. Park, B.K., et al., *Drug bioactivation and protein adduct formation in the pathogenesis of drug-induced toxicity*. Chem Biol Interact, 2011. **192**(1-2): p. 30-6.
83. Hanzlik, R.P., Y.M. Koen, and J. Fang, *Bioinformatic analysis of 302 reactive metabolite target proteins. Which ones are important for cell death?* Toxicol Sci, 2013. **135**(2): p. 390-401.
84. Hanzlik, R.P., et al., *The reactive metabolite target protein database (TPDB)--a web-accessible resource*. BMC Bioinformatics, 2007. **8**: p. 95.
85. Ju, C. and T. Reilly, *Role of immune reactions in drug-induced liver injury (DILI)*. Drug Metab Rev, 2012. **44**(1): p. 107-15.
86. Williams, C.D. and H. Jaeschke, *Role of innate and adaptive immunity during drug-induced liver injury*. Toxicol Res (Camb), 2012. **1**: p. 161-170.
87. Keogh, J.P., *Membrane transporters in drug development*. Adv Pharmacol, 2012. **63**: p. 1-42.
88. Vinken, M., et al., *Development of an adverse outcome pathway from drug-mediated bile salt export pump inhibition to cholestatic liver injury*. Toxicol Sci, 2013. **136**(1): p. 97-106.
89. Kis, E., et al., *BSEP inhibition: in vitro screens to assess cholestatic potential of drugs*. Toxicol In Vitro, 2012. **26**(8): p. 1294-9.
90. Dawson, P.A., T. Lan, and A. Rao, *Bile acid transporters*. J Lipid Res, 2009. **50**(12): p. 2340-57.
91. Reshetnyak, V.I., *Physiological and molecular biochemical mechanisms of bile formation*. World J Gastroenterol, 2013. **19**(42): p. 7341-60.
92. Anwer, M.S., *Cellular regulation of hepatic bile acid transport in health and cholestasis*. Hepatology, 2004. **39**(3): p. 581-90.
93. Arrese, M. and M. Trauner, *Molecular aspects of bile formation and cholestasis*. Trends Mol Med, 2003. **9**(12): p. 558-64.
94. Giacomini, K.M., et al., *Membrane transporters in drug development*. Nat Rev Drug Discov, 2010. **9**(3): p. 215-36.
95. Hillgren, K.M., et al., *Emerging transporters of clinical importance: an update from the International Transporter Consortium*. Clin Pharmacol Ther, 2013. **94**(1): p. 52-63.
96. Palmieri, F., *The mitochondrial transporter family SLC25: identification, properties and physiopathology*. Mol Aspects Med, 2013. **34**(2-3): p. 465-84.
97. Zutz, A., et al., *Mitochondrial ABC proteins in health and disease*. Biochim Biophys Acta, 2009. **1787**(6): p. 681-90.
98. Botts, S., et al., *Introduction to hepatic drug metabolizing enzyme induction in drug safety evaluation studies*. Toxicol Pathol, 2010. **38**(5): p. 796-8.
99. Zollner, G., M. Wagner, and M. Trauner, *Nuclear receptors as drug targets in cholestasis and drug-induced hepatotoxicity*. Pharmacol Ther, 2010. **126**(3): p. 228-43.
100. Claudel, T., et al., *Role of nuclear receptors for bile acid metabolism, bile secretion, cholestasis, and gallstone disease*. Biochim Biophys Acta, 2011. **1812**(8): p. 867-78.
101. Lopez-Velazquez, J.A., et al., *Nuclear receptors in nonalcoholic Fatty liver disease*. J Lipids, 2012. **2012**: p. 139875.

102. Chu, V., et al., *In vitro and in vivo induction of cytochrome p450: a survey of the current practices and recommendations: a pharmaceutical research and manufacturers of america perspective*. Drug Metab Dispos, 2009. **37**(7): p. 1339-54.
103. LiverTox, *Angiotensin-converting enzyme (ACE) inhibitors*. 2014.
104. Roberts, S.M., R.P. DeMott, and R.C. James, *Adrenergic modulation of hepatotoxicity*. Drug Metab Rev, 1997. **29**(1-2): p. 329-53.
105. Ma, W.L., et al., *Androgen receptor roles in hepatocellular carcinoma, fatty liver, cirrhosis and hepatitis*. Endocr Relat Cancer, 2014. **21**(3): p. R165-82.
106. LiverTox. *Beta-2 Adrenergic Agonists*. 2014; Available from: <http://livertox.nih.gov/Beta2AdrenergicAgonists.htm>.
107. LiverTox, *Beta-Adrenergic Receptor Antagonists (Beta-Blockers)*. 2014.
108. LiverTox. *Carbonic Anhydrase Inhibitor Diuretics*. 2014; Available from: <http://livertox.nih.gov/CarbonicAnhydraseInhibitorDiuretics.htm>.
109. PatientPlus. *Diuretics*. 2014; Available from: <http://www.patient.co.uk/doctor/diuretics>.
110. Clemens, M.G., *Nitric oxide in the liver*, in *The Liver: Biology and Pathobiology*, I.M. Arias, et al., Editors. 2001, Lippincott Williams and Wilkins. p. 555-564.
111. Unzueta, A. and H.E. Vargas, *Nonsteroidal anti-inflammatory drug-induced hepatotoxicity*. Clin Liver Dis, 2013. **17**(4): p. 643-56, ix.
112. LiverTox. *Apomorphine*. 2014; Available from: <http://livertox.nlm.nih.gov/Apomorphine.htm>.
113. Xu, J.W., et al., *Effects of estradiol on liver estrogen receptor-alpha and its mRNA expression in hepatic fibrosis in rats*. World J Gastroenterol, 2004. **10**(2): p. 250-4.
114. Macias Saint-Gerons, D., et al., *[Hepatotoxicity in patients treated with endothelin receptor antagonists: systematic review and meta-analysis of randomized clinical trials]*. Med Clin (Barc), 2014. **142**(8): p. 333-42.
115. LiverTox. *H2 Receptor Blockers*. 2014; Available from: <http://livertox.nlm.nih.gov/H2ReceptorBlockers.htm>.
116. Da Prada, M., et al., *Preclinical profiles of the novel reversible MAO-A inhibitors, moclobemide and brofaromine, in comparison with irreversible MAO inhibitors*. J Neural Transm Suppl, 1989. **28**: p. 5-20.
117. LiverTox. *Erythromycin*. 2014; Available from: <http://livertox.nih.gov/Erythromycin.htm>.
118. Zuckerman, J.M., *Macrolides and ketolides: azithromycin, clarithromycin, telithromycin*. Infect Dis Clin North Am, 2004. **18**(3): p. 621-49, xi.
119. Kim, S.W., et al., *Biphasic changes of epithelial sodium channel abundance and trafficking in common bile duct ligation-induced liver cirrhosis*. Kidney Int, 2006. **69**(1): p. 89-98.
120. LiverTox, *Dronedarone*. 2014.
121. LiverTox, *Memantine*. 2014.
122. Ulrich, R.G., *Idiosyncratic toxicity: a convergence of risk factors*. Annu Rev Med, 2007. **58**: p. 17-34.
123. Parkinson, A. and B.W. Ogilvie, *Biotransformation of Xenobiotics*, in *Casarett & Doull's Toxicology: The Basic Science of Poisons*, C. Klaassen, Editor. 2007, McGraw-Hill Professional. p. 161-304.
124. Affymetrix. *DMET™ Plus Solution*. 2014; Available from: http://www.affymetrix.com/catalog/131412/AFFY/DMET-Plus-Solution - 1_1.
125. Cerep, *ADME-Tox*. 2014.
126. Cyprotex, *In vitro ADME and PK*. 2014.
127. ChEMBL. *ADME SARfari*. 2014; Available from: <https://http://www.ebi.ac.uk/chembl/admesarfari>.
128. Cotter, D., et al., *MitoProteome: mitochondrial protein sequence database and annotation system*. Nucleic Acids Res, 2004. **32**(Database issue): p. D463-7.
129. Pagliarini, D.J., et al., *A mitochondrial protein compendium elucidates complex I disease biology*. Cell, 2008. **134**(1): p. 112-23.

130. Smith, A.C. and A.J. Robinson, *MitoMiner, an integrated database for the storage and analysis of mitochondrial proteomics data*. Mol Cell Proteomics, 2009. **8**(6): p. 1324-37.
131. *ConsensusPathDB*. 2014; Available from: <http://consensuspathdb.org/>.
132. Logue, J.S. and D.K. Morrison, *Complexity in the signaling network: insights from the use of targeted inhibitors in cancer therapy*. Genes Dev, 2012. **26**(7): p. 641-50.
133. Rognstad, R., *Rate-limiting steps in metabolic pathways*. J Biol Chem, 1979. **254**(6): p. 1875-8.
134. Thiele, I., et al., *A community-driven global reconstruction of human metabolism*. Nat Biotechnol, 2013. **31**(5): p. 419-25.
135. *DrugMatrix*. 2014; Available from: <https://ntp.niehs.nih.gov/drugmatrix/index.html>.
136. Kiyosawa, N., et al., *Gene set-level network analysis using a toxicogenomics database*. Genomics, 2010. **96**(1): p. 39-49.
137. *CTD: Comparative Toxicogenomics Database*. 2014; Available from: <http://ctdbase.org/>.
138. Davis, A.P., et al., *A CTD-Pfizer collaboration: manual curation of 88,000 scientific articles text mined for drug-disease and drug-phenotype interactions*. Database (Oxford), 2013. **2013**: p. bat080.
139. *PharmGKB*. 2014; Available from: <http://www.pharmgkb.org/>.
140. Agrawal, Y.P. and H. Rennert, *Pharmacogenomics and the future of toxicology testing*. Clin Lab Med, 2012. **32**(3): p. 509-23.
141. Andrade, R.J., et al., *Pharmacogenomics in drug induced liver injury*. Curr Drug Metab, 2009. **10**(9): p. 956-70.
142. Roden, D.M., et al., *Cardiovascular pharmacogenomics*. Circ Res, 2011. **109**(7): p. 807-20.