Development of informatic tools for extracting biomedical data from open and proprietary data sources with predictive purposes

Oriol López Massaguer

TESI DOCTORAL UPF / 2017

DIRECTOR DE LA TESI

Dr. Manuel Pastor Maeso

DEPARTAMENT CEXS

Universitat
Pompeu Fabra
Barcelona

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking Open PHACTS Project, grant agreement number 115191 and eTOX project, grant agreement number 115002, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contributions.

A mis padres

A la rosa sin porqué

La rosa es sin porqué, florece porque florece.

Die Rose ist ohne warum; sie blühet weil sie blühet.

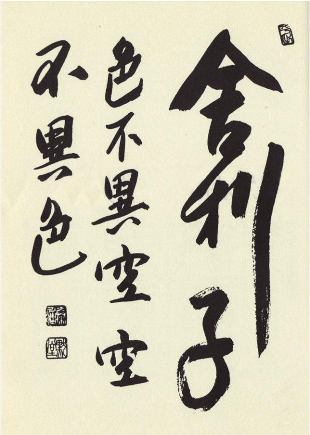
Cherubinischer Wandersmann

Wir fühlen, daß selbst, wenn alle *möglichen* wissenschaftlichen Fragen beantwortet sind, unsere Lebensprobleme noch gar nicht berührt sind.
Freilich bleibt dann eben keine Frage mehr; und eben dies ist die Antwort.

We feel that even if all possible scientific questions be answered, the problems of life have still not been touched at all.
Of course there is then no question left, and just this is the answer.[1]

Sentimos que aun cuando todas las *posibles* cuestiones científicas hayan recibido respuesta, nuestros problemas vitales todavía no se han rozado en lo más mínimo.
Por supuesto que entonces ya no queda pregunta alguna;
y esto es precisamente la respuesta.[2]

Ludwig Wittgenstein, Tractatus logico-philosophicus, 6.52, 1922



¡Oh, Sariputra,
la forma es vacío,
el vacío es una forma.

Maha Prajna Paramita Hridaya Sutra
Taisen Deshimaru roshi calligraphy

[1] Translation C. K. Ogden (1922) with F. P. Ramsey, G. E. Moore and L. Wittgenstein

[2] Traducción Jacobo Muñoz e Isidoro Reguera (1973). Alianza Editorial.

Agradecimientos

A Manuel Pastor porque me dio la oportunidad de realizar este doctorado. A Ferran Sanz por su liderazgo en el GRIB.

A Nuria Centeno porque no me disuadió de hacer el Master en Bioinformática.

A Pau Carrió por su apoyo y compañerismo durante estos años.

A Martina, Chus, Maria y Carina del staff del GRIB por su profesionalidad, disponibilidad y amabilidad.

A Miguel Ángel y Alfons de IT por soportar mis continuas peticiones de soporte con amabilidad y profesionalidad.

A Alexander Amberg y Lennart T. Anger de SANOFI por su colaboración en la investigación que condujo a esta tesis doctoral en el marco del proyecto eTOX.

A Philippe Marc y Carlo Ravagli de Novartis por su colaboración que condujo a la idea de explotar la ontologías en el marco del proyecto eTOX.

A Antonia por sus bollos con chocolate que alegraron cada mañana de este PhD.

Summary

We developed new software tools to obtain information from public and private data sources to develop *in silico* toxicity models. The first of these tools is Collector, an Open Source application that generates “QSAR-ready” series of compounds annotated with bioactivities, extracting the data from the Open PHACTS platform using semantic web technologies. Collector was applied in the framework of the eTOX project to develop predictive models for toxicity endpoints.

Additionally, we conceived, designed, implemented and tested a method to derive toxicity scorings suitable for predictive modelling starting from *in vivo* preclinical repeated-dose studies generated by the pharmaceutical industry.

This approach was tested by generating scorings for three hepatotoxicity endpoints: ‘degenerative lesions’, ‘inflammatory liver changes’ and ‘non-neoplasic proliferative lesions’. The suitability of these scores was tested by comparing them with experimentally obtained point of departure doses as well as by developing tentative QSAR models, obtaining acceptable results.

Our method relies on ontology-based inference to extract information from our ontology annotated data stored in a relational database.

Our method, as a whole, can be applied to other preclinical toxicity databases to generate toxicity scorings. Moreover, the

ontology-based inference method on its own is applicable to any relational databases annotated with ontologies.

Resum

Hem desenvolupat noves eines de software per tal d’obtenir informació de fonts publiques i privades per tal de desenvolupar models de toxicitat *in silico*. La primera eina es Collector, una aplicació de programari lliure que genera series de compostos preparats per fer modelat QSAR anotats amb bioactivitats extretes de la plataforma Open PHACTS usant tecnologies de la web semàntica. Collector ha estat utilitzada dins el projecte eTOX per desenvolupar models predictius sobre *endpoints* de toxicitat.

Addicionalment hem concebut, desenvolupat i implementat un mètode per derivar *scorings* de toxicitat apropiats per modelatge predictiu que utilitza les dades obtingudes de informes d’estudis amb dosis repetides *in vivo* de la industria farmacèutica.

El nostre mètode ha estat testejant aplicant-lo al modelat de hepatotoxicitat obtenint les dades corresponents per 3 endpoints: ‘degenerative lesions’, ‘inflammatory liver changes’ and ‘non-neoplasic proliferative lesions’. S’ha validat la idoneïtat d’aquestes dades obtingudes comparant-les amb els valors de *point of departure* obtinguts experimentalment i també desenvolupant models QSAR de prova obtenint resultats acceptables.

El nostre mètode es basa en la inferència basada en ontologies per extreure informació de la nostra base de dades on tenim dades anotades basades en ontologies.

El nostre mètode també es pot aplicar a altres bases de dades amb informació preclínica per generar *scorings* de toxicitat. Addicionalment el nostre mètode d'inferència basat en ontologies es pot aplicar a d'altre bases de dades relacionals anotades amb ontologies

Preface

The present work describes the development of new tools for obtaining information from diverse sources and building better predictive models for *in silico* toxicology, with the potential of reducing the needs of carrying out *in vivo* toxicity testing. This work was developed in the framework of the eTOX IMI/JU project, in close collaboration with pharmaceutical companies. We believe that our methods and tools will be applied in the future and will have a real impact in the toxicity methods used by the pharmaceutical industry, contributing to a more efficient safety testing and enabling 3R (reduce, refine and replace) policies for animal testing.

In particular, the approach has already been tested by the company SANOFI to model hepatotoxicity endpoints and this collaboration will be extended in the future.

Table of contents

Figure index

1 INTRODUCTION

1.1 Overview of drug discovery and development process

1.1.1 STEPS OF THE PROCESS

The drug discovery and development process aims to develop therapeutically useful and safe drugs. This long and costly process spans different phases (see Figure 1: Drug discovery and development process). The most recent estimations are that average R&D cost required to bring a new, FDA[3]-approved medicine to patients is of $2600 million (2013 dollars) and it takes 10 to 15 years [1]:

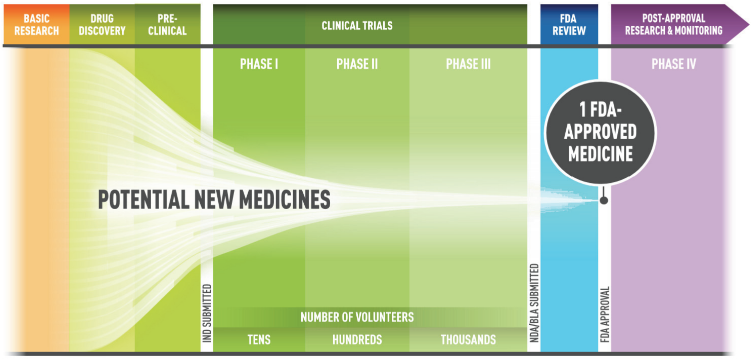


Figure 1: Drug discovery and development process[4] [2]

[3] US Food and Drug Administration. https://www.fda.gov/

[4] Key: IND: Investigational New Drug Application, NDA: New Drug Application BLA: Biologics License Application

The following description is a concise overview of a typicall drug discovery and development process, but it can vary depending on the pharma companies involved and the feedback from regulators[5] during the drug approval. The process[3] involves several steps:

- Target identification and validation: researchers identify biological targets for a potential disease. A target is a molecular structure in the body that, when it interacts with a drug, produces a therapeutic effect (e.g., treatment or prevention of a disease).

- Lead identification: identitication of compounds that show pharmacological activity against the biological target. This is usually done by using high throughput screening techniques.

- Lead optimization: Lead compounds are “optimized” or altered to improve efficacy and safety. By altering structurally the compounds, researchers can modulate their properties. The compounds identified as the “best” one, meeting the defined criteria become a “drug candidate” that undergoes further testing.

[5] FDA: https://www.fda.gov, EMA: http://www.ema.europa.eu,
PMDA: https://www.pmda.go.jp

- Preclinical safety testing: Before testing a drug in humans, researchers must find out whether it has the potential to cause serious harm, also called toxicity. Both *in vitro* and *in vivo* studies are conducted. These studies must provide detailed information on dosing and toxicity levels. After preclinical testing, researchers review their findings and decide whether the drug should be tested in human.

- Investigational new drug application: Before any clinical trial the drug developer must file an application to the regulator that includes the results of the preclinical tests, the drug structure and properties and the possible side effects detected in preclinical tests. It also contains a detailed plan of the clinical testing.

- Clinical tests: The clinical research phase involves testing the drug in humans to determine safety and efficacy. It is divided into several sub-phases:

Phase	Goal	Participants	Duration	Next phase % move
Phase 1	Safety Dosage	20 to 100 healthy volunteers	Months	70 %
Phase 2	Efficacy and side effects	Up several hundred people with disease / condition	Months to 2 years	33 %
Phase 3	Efficacy and monitoring of adverse reactions	300 to 3000 people with disease / condition	1 to 4 years	25-30%
Phase 4	Safety and efficacy	Several thousand people with disease / condition	Post-market approval	NA

- Regulator review: Once the drug developer has obtained enough scientific evidence that the candidate is safe and effective for its intended use from the previous studies, the company can file an application to market that will be reviewed by the regulator. The regulator reviews the evidence, which must be provided in standard formats and judges if actually the drug is safe and effective for its intended use. In case of any doubt, the regulators can request further testing to be performed, and if the developers fails to fulfil their strict requirements, the drug candidate is rejected.

- Post-Market Safety Monitoring: After market approval the regulators monitor the safety of all drugs during the product lifetime. Regulators reviews reports of problems and may decide to add cautions to the dosage or usage information, as well as other measures for more serious issues, like deciding the withdrawal of the product from the market

Although the previous steps are described in a sequential way there is a certain back and forth between certain steps. The purpose of the description is only to provide a high level overview of the process and can vary in different pharma companies or depending on the targeted disease.

All in all, the goal of the process is not only to obtain new compounds but to guarantee their therapeutic efficacy and safety in humans.

1.1.2 Challenges and Shortcomings

In the latest 60 years it was observed a steady decrease in the drug research and development productivity, as illustrated in Figure 2 [4]:

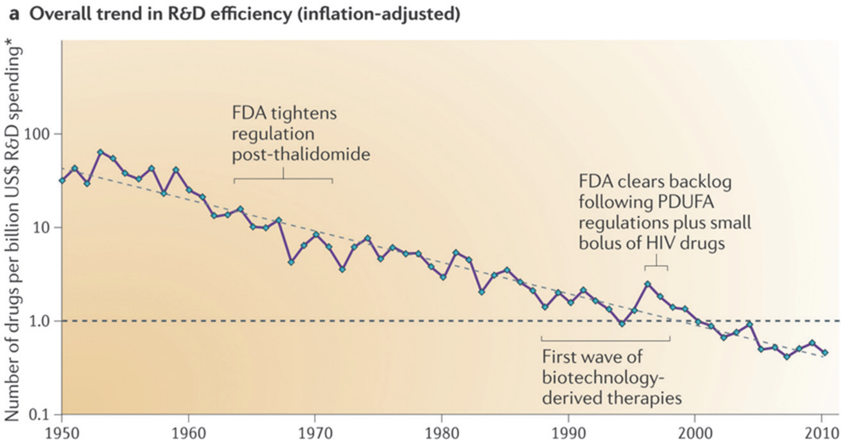


Figure 2: Decrease drug R&D efficiency

The above chart shows that the number of new drugs approved by the FDA per billion US dollars (inflation-adjusted) spent in research and development (R&D) has halved roughly every 9 years. This tendency has been coined as “Eroom’s law”

According to this tendency, the cost of developing a new drug has increased steadily (data until 2013) [1]:

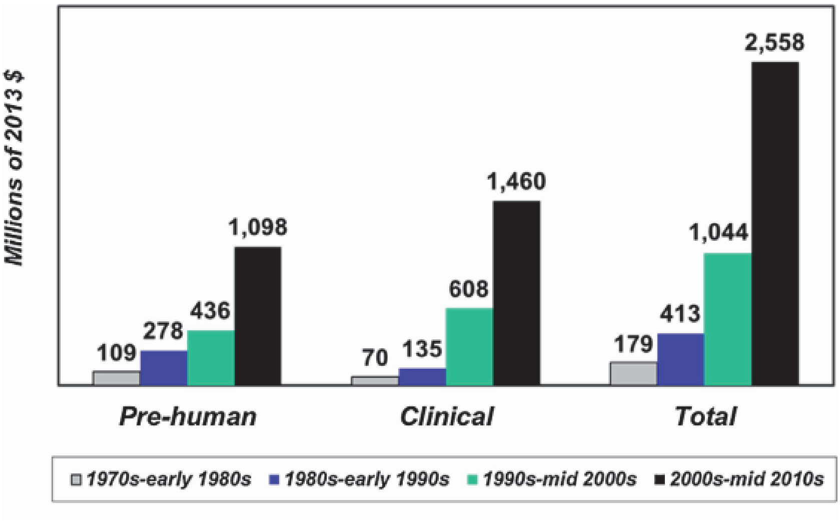


Figure 3: Trends in total cost per new approved drug

In recent times (since 2008) we observed a certain improvement of this trend [5]. However, productivity depends both on returns and spending and we can see a stabilization of R&D budget since 2008:

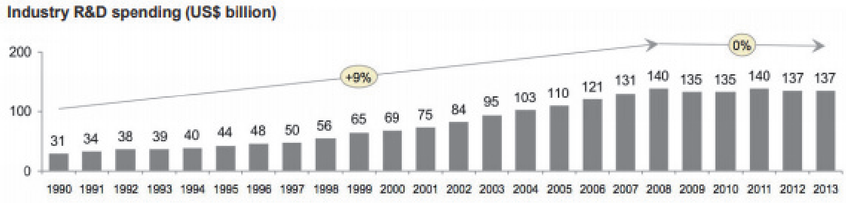


Figure 4: Pharma industry R&D spending

So, the slight increase in productivity could be attributed to the stabilization of the R&D budgets and it is not clear if this budget stabilization will impact long-term productivity.

There is no general consensus about the causes of this decline in productivity (see [4] for a review). But, most likely, part of the causes are related to the way that basic research is conducted and the way pharma companies drive the process. Scanell et al [4] identified two main potential problems:

- The idea that high-affinity binding to a single biological target linked to a disease will lead to medical benefit in humans. The evidence of this hypothesis is weaker than previously thought. Promiscuous binding of small molecules to several targets has an impact on the efficacy and safety which is not fully understood and which will require new research tools and paradigms.
- The second potential problem is related to the nature of chemical space and the recent shift from iterative medicinal chemistry coupled with parallel assays (pre-1990s) to serial filtering that begins with HTS[6] of a static compound library against a target.

In short, the shift in Pharmaceutical R&D from more complex and labour intensive animal model testing to higher

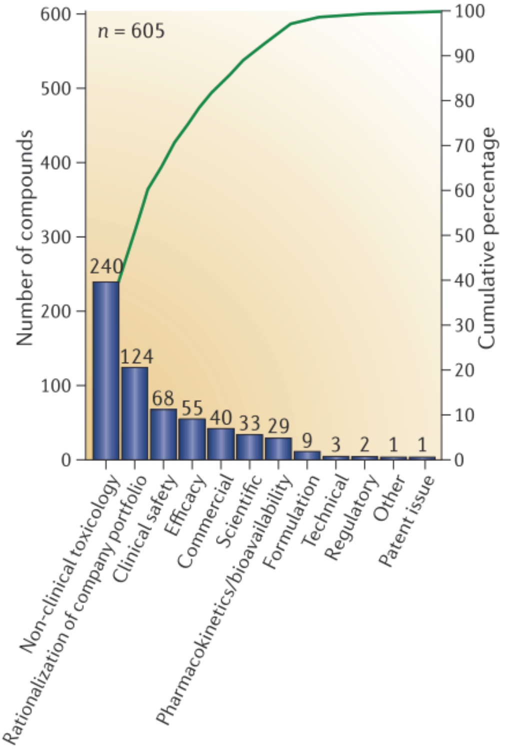
[6] High Troughput Screening

throughput but simpler methods like HTS [4] seems to have degraded productivity. It looks reasonable to improve current methods to tackle the complexity of the biological and chemical space avoiding brute-force approaches.

1.1.3 SAFETY CONCERNS

Besides the need to improving the productivity of the process there is an urging need to produce safer drugs for several reasons: ethical, human safety, etc.

Approximately 35% of all drug development projects fail as a result of toxicity detected during preclinical safety studies [2] (which impacts in productivity). In other cases, drug failures occur in clinical testing phases or even in post-market phases. In the later cases (apart from the impact in productivity and on the value of the company in the stock market) the failure has significant ethical implications. For these reasons, reducing the attrition rates of the pharma companies is a major need. One of the primary causes of attrition are toxicity concerns found at either preclinical or clinical stages [6]. The next chart shows attrition causes in a dataset compiled from four large Pharma companies:



Despite the huge investment [7] in *in vitro* toxicology screening made by the industry, the attrition rates are similar in the two time periods studied by the authors [6]. However, because the underlying mechanism of toxicity-related attrition was not included, it is also possible that although the incidence of toxicology failures has remained the same the cause of the failures may have changed over the years.

So, in spite of the improvements in the detection of frequent and dangerous adverse events like genotoxicity or hERG inhibition [6], we are far from having a perfect safety testing.

It is important to stress that it is critical to detect potential toxicity as early as possible both for economical (the later the failure the more money invested) and ethical reasons (to avoid unnecessarily animal testing or withdrawals from the market due to safety concerns)

The withdrawal of marketed drugs (such as Ximelagatran [8] or Vioxx[9]) were due, in most cases, to toxicities involving mechanisms harder to predict in preclinical studies (from *in vitro* data and from model species) or even in clinical studies. For example, in the case of idiosyncratic toxicity [10][11] only a small number of patients developed severe toxic effects, making it very hard to detect pre-marketing.

Another parallel issue is the increasing social pressure to reduce animal testing due to ethical concerns [12]. For example the REACH[7] European regulation framework [13] derived in the banning of cosmetics testing on animals [14].

[7] Registration, Evaluation, Authorisation and Restriction of Chemicals

1.1.4 Paradigm shift in risk assessment

Existing toxicological tests are based on the identification of an adverse effect at a given dose, which can be used to define a point of departure [15][16] (PoD) for subsequent assessment of risk and regulatory decision-making. PoDs were often calculated as the highest dose at which no adverse effect was detected (NOAEL, no-observed-adverse-effect level)[8] or the lowest dose at which an adverse effect was observed (LOAEL, lowest-observed-adverse-effect level)[9] values.

For drugs, the dose at which any relevant adverse effect is detected should be much higher than the therapeutic dose for this drug to be considered amenable of being marketed.

[8] Greatest concentration or amount of a substance, found by experiment or observation, which causes no detectable adverse alteration of morphology, functional capacity, growth, development, or life span of the target organism under defined conditions of exposure.

[9] Lowest concentration or amount of a substance (dose), found by experiment or observation, which causes an adverse effect on morphology, functional capacity, growth, development, or life span of a target organism distinguishable from normal (control) organisms of the same species and strain under defined conditions of exposure.

In the pharmaceutical area, PoD are often derived from the observation of apical endpoints in repeated-dose testing in animals, according to clearly defined regulatory guidelines. Their determination involves expert toxicologist judgment that decides if the effect observed is related to the tested compound, is adverse and clinically relevant.

However, recently it has been proposed a paradigm shift [17] in toxicity assessment from the observation of classical apical endpoints towards an approach based on a deeper mechanistic understanding of toxicity.

This new paradigm appears from the need of improving the shortcomings of our current testing paradigm. Animal testing although it is an essential part of toxicity testing has been challenged due to limited translatability to humans [18].

A major component of this change in strategy is the development of a framework allowing to represent existing knowledge about toxicity mechanisms at multiple levels, in the form of a causal cascade of events, the so-called Adverse Outcome Pathway (AOPs) [19] [20].

In short AOPs can be defined as:

“An AOP is an analytical construct that describes a sequential chain of causally linked events at different levels of biological organisation that lead to an adverse health or ecotoxicological effect. AOPs are the central element of a toxicological knowledge framework being built to support chemical risk assessment based on mechanistic reasoning." [21].

The AOPs are illustrated in figure 5:

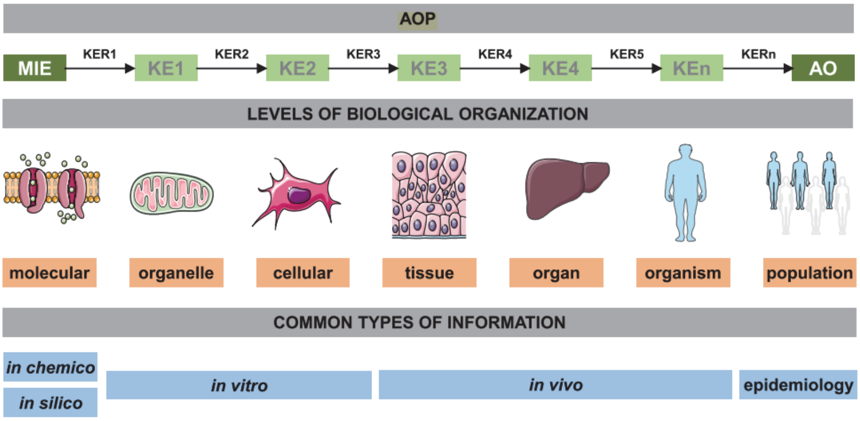


Figure 5: AOP structure[20]

AOPs connect a Molecule Initiating Event (MIE) to a sequence of Key Events (KE) connected by Key Event Relationships (KER),producing an Adverse Outcome (AO). The AOP models the causal chain of events reflecting the modularity of complex process and it spans from the molecular level to the organism and population levels.

AOPs were intended specifically to support regulatory decision-making based on the desire to make effective use of mechanistic data, but they also can be applied to characterize the identify chemical categories based on biological responses [20]. Furthermore, once the chemical category is fully established, it can be used for data gap filling strategies, such as read-across techniques that apply relevant information from analogous substances to predict the toxicological properties of a target substance.

This paradigm shift was embraced by a series of initiatives and projects. We will briefly mention some of these projects: Tox21 [22] project by the EPA[10]: By using a high-throughput robotic screening system housed at NCATS, researchers are testing 10,000 environmental chemicals (called the Tox21 10K library) for their potential to disrupt biological pathways that may result in toxicity. Screening results help the researchers to prioritize chemicals for further in-depth investigation. The eTOX project [23][7] aims to collect existing information from repeated-dose toxicity (RDT) studies carried out in the pharmaceutical industry and to exploit this information, both directly and by developing models able to predict *in vivo* toxicity endpoints.

Not directly related with toxicity, other initiatives such as the Open PHACTS project [24] have contributed to improve existing methodologies for knowledge management in the drug development and discovery process.

[10] U.S. Environmental Protection Agency. https://www.epa.gov/

1.2 Computational methods in toxicology

1.2.1 OVERVIEW AND MOTIVATION

Animal testing is used routinely for toxicity testing and, as we mentioned, is still a critically important part of the regulatory studies required to obtain the drug approval. *In vitro* testing is increasingly used due to advances in HTS technologies and their lower cost, time and compound consumption. They are particularly used in discovery phases, for the prioritization of candidates and the early identification of safety concerns, also because regulatory rules do not constrain its use.

An interesting alternative is *in silico* toxicology. This technique uses computational tools to assess toxicity [25].

The goal of computational methods is to help *in vitro* and *in vivo* testing to reduce animal testing, lower costs of toxicity testing and improve safety assessment.

The advantages of *in silico* toxicology are lower cost and less time and ethical concerns since they do not require the use of experimentation animals. Additionally, computational methods have the advantage of avoiding the need of consuming valuable compounds, and indeed, testing can be applied even before they are synthesized.

1.2.2 COMPUTATIONAL TOOLS

In silico toxicology uses a broad range of tools to fulfil their goals: databases (storing molecular and biological data), molecular descriptors, molecular dynamics simulations, system biology simulations, modelling methods and software to develop such models and expert systems.

In this work, we will focus our attention on databases used in toxicity modelling and how to extract information from these databases for modelling purposes.

1.2.3 TOXICITY MODELLING

Toxicity modelling usually includes five major steps: gathering data that links chemicals with biological responses, computing molecular descriptors of the chemicals, generating a prediction model, evaluate the accuracy of the model and interpreting the model.

For the model development, there are different strategies such as (adapted from [25]):

- Structural Alerts: Try to identify chemical sub-structures that are associated with toxicity
- Read-across: predict the toxicity of chemicals using toxicity data from similar chemicals

- Dose-response and time-response models: Try to link the dose (or time) with the biological effect (toxic)

- Pharmacokinetic (PK) and Pharmacodynamic (PD): These models try to model the concentration/time of the drug in the organ/tissue relate to a toxic response

- Uncertainty factor models (UFs): UFs are used to assess the risk of chemical exposure or daily intake of chemicals. They were routinely used to extrapolate between species based on the PoD obtained from animal testing (NOAEL, LOAEL, ..)

- QSAR (Quantitative Structure-Activity Relationship) models: they use molecular descriptors to link chemical structure to a biological response, in this case a toxic response.

The work of this thesis is focused on data gathering to provide data series suitable for developing QSAR models for toxicity prediction. We developed specialized software tools for gathering *in vitro* data from public resources (see 1.3) and proprietary preclinical data (see 1.4). Also, we used our group expertise and tools (eTOXlab [26]) to develop and test such models (see 3.4.1 and 3.4.2)

1.2.4 THE ROLE OF DATA IN MODELLING

Data is probably the most critical component in the development of any computational tool able to predict drug toxicity. Ultimately, any such method works starting from this data, identifying patterns associated with the toxic properties, and therefore, the quality of the data imposes unsurpassable limits to the quality of the results obtained.

For example, the data needed to develop QSAR models include [27]

- Chemical data: includes chemical and chemical properties as well as computed chemical descriptors.
- Effect or property that we want to predict: It includes biological properties and other properties that can be used to (in our case) measure toxicity.

Data for modelling can be obtained from different sources:

- Public data sources: It comprises public bioactivity (*in vitro* or *in vivo*) databases such as ChEMBL[28], DrugBank[29] , PubChem[30], ToxCast[31], ToxRefDb[32], patent databases, public literature.
- Proprietary databases: It comprises databases from private providers and in-house databases from pharma companies.

But it is difficult to obtain these data in a consistent and integrated way.The biggest challenge is to obtain data of enough quality with respect to the following criteria:

- Heterogeneity, data generated with different purposes and data formats.
- Integration, need for data normalization and identity mapping .
- Quality, varying degree from the experimental or primary sources, and varying quality in storage and structuring level.
- Curation, degree of automatic or human based data curation.
- Comparability, data obtained in different experimental conditions for example.

Another important point in the area of drug discovery and development is that some of the data is stored within pharma companies data-silos and/or is proprietary or confidential, such as the chemical structures of drug in development and preclinical testing data.

There have been recent effort to overcome this situation both in the data integration and the access to private data silos in several public-private partnerships of two IMI projects Open PHACTS [24] and eTOX [7].

1.3 Extraction of modelling data from open data sources

1.3.1 DATA INTEGRATION DIFFICULTIES

In biomedical research, there is an urgent need to improve the way we manage and obtain data. The different providers and institutions involved (academia, funding agencies, data providers) are increasingly concerned about the way the data is managed. Good data management includes [33] proper data collection, annotation and archival. But it is increasingly important that data was machine-actionable. This means that data have to follow certain principles that make data easily tractable by machines. The FAIR[11] principles have been defined and they establish a set of guidelines of data stewardship (See [33] Box 2 for more details):

- Findable: data and metadata are assigned a unique global identifier and have to be available in a searchable resource

- Accessible: data and metadata have to be accessible through his unique identifier using an open and free protocol

[11] Findable Accessible Interoperable Re-usable.

- Interoperable: data and metadata have to be represented using a language for knowledge representation. The vocabularies used have to adhere FAIR principles.

- Reusable: data and metadata must be described with accurate and relevant attributes. They also have to include provenance and a clear license.

Recently, the IMI project Open PHACTS [34] has built a platform for researchers to access and query publicly-available pharmacological data. The Open PHACTS platform delivers data in a single, integrated and open infrastructure. It adheres the FAIR principles [33]. It is not specifically focused in toxicity but the data therein is very useful for toxicity modelling.

Open PHACTS system is based on semantic web technologies to improve the process of drug discovery by:

- Developing a set of robust standards to enable integration between data sources via semantic technologies and implementing them in a semantic integration hub ("Open Pharmacological Space" or OPS)

- Integrating the information and transforming it into a common form of representation using semantic web [35] [36] technologies

- Developing several tools (known as exemplar applications or eApps) allowing easy access to the information and answering relevant research questions.

1.3.2 Semantic web and Open PHACTS

The Open PHACTS uses of semantic web technologies [35,36] (see Figure 6: What is the semantic web?) as a strategy and technology to solve the issues, mentioned above, related to data integration explained. This technology has major advantages:

- Uniform data model: based on Resource Description Framework [37], RDF Schema [38] and Web Ontology Language [39]

- Identity: a clear method to guarantee the identity of the entities modelled via the URI concept [40]

- Logical interpretation: a clear unambiguous semantic interpretation of data based on Descriptions logics [41].

- Mapping mechanisms: by using standard ontologies such as SKOS [42] and VoiD [43]. These ontologies will be used to map the identity of objects from different data sources.

- Open World Assumption (OWA): that allows working with incomplete or lacking information [44]. The open world assumption states that if some fact is unknown cannot be supposed to be false. On the contrary the closed world assumption (CWA) states that if some fact is unknown is assumed to be false. The CWA is assumed in systems such as relational databases or IT systems that have a complete control of information. This assumption cannot be hold on semantic web systems.

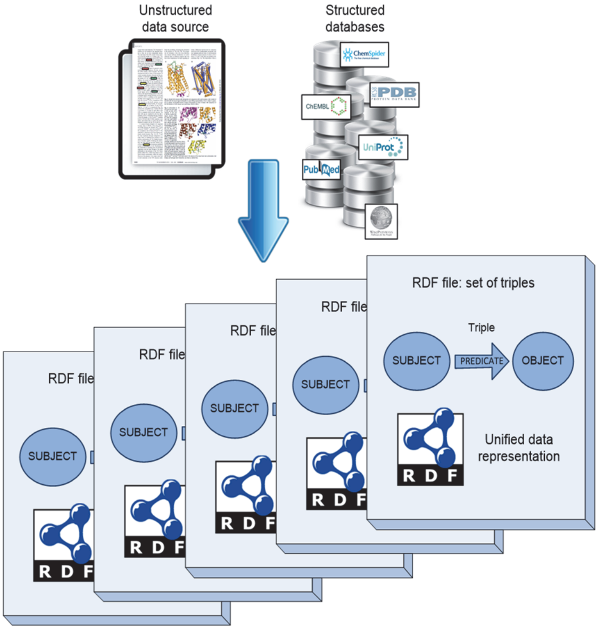


Figure 6: What is the semantic web?

- An unified and clear licensing model for the data extracted [45]

1.3.3 Open PHACTS platform

Open PHACTS provides a data integration platform focused on drug discovery research:

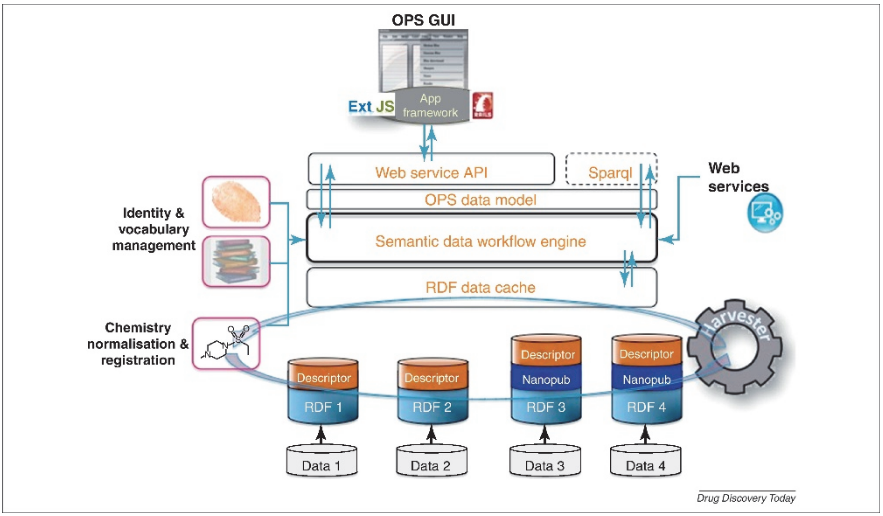


Figure 7: Open PHACTS platform [46]

The Open PHACTS platform cover different needs:

- Data integration: it integrates into a single platform different datasets relevant for drug discovery research[12]

- Data Access: The data in the platform is accessible via a unified API[13]

- Identify and name resolution: It provides tools for mapping the identity of the different concepts and entities of the domain via the identity resolution and searching for entities through name resolution provided by the Open PHACTS API[14]

- Simplified API access: although semantic web data is usually accessed by using SPARQL [47], it is a complex language and it is easy to write poorly

[12] https://www.openphacts.org/2/sci/data.html

[13] https://dev.openphacts.org/docs/2.2

[14] https://dev.openphacts.org/

performant queries. So Open PHACTS decided to encapsulate data access through a web API.

1.3.4 Exploiting the OPS platform

In the life of the project, several applications were developed. The aim was mainly to test and to demonstrate the capabilities of the OPS, and they were designed to cover more specific needs in the areas of drug development. The main eApps were:

eApps	Description
Pharmatrek [48]	PharmaTrek proposes new mechanisms to navigate the pharmacological space in an interactive and flexible way. PharmaTrek is an integrative and interactive web application that allows the scientist to extract new knowledge from the Open PHACTS platform. The main goal is to provide visual tools that allow the user to define custom questions around the biological activity between drugs and targets
GARField [49]	GARField is a tool for exploring the pharmacological space of small molecules. The user can search a compound from name, synonym, SMILES or structure drawing and retrieve target information associated with the compound query. Similarly, for a

	given protein, compounds with bioactivity data can be shown.
ChemBioNavigator[50]	The ChemBioNavigator (CBN) allows navigation of the interface of chemical and biological data, tailored for applications in pharmaceutical research. CBN lets you easily browse through sets of compounds: different sorting and plotting options offer a quick and intuitive overview of the physicochemical characteristics.
Target Dosier[51]	The Target Dossier uses the Open PHACTS platform to build comprehensive views of pharmacologically relevant targets, to answer questions regarding druggability, tissue expression profiles and implications in pathways, disease states and physiological mechanisms. The Target Dossier will provide a decision support platform for target selection and progression.
Explorer [52]	The Open PHACTS Explorer provides a basic, all-purpose tool for browsing and querying all of the pharmacological and physicochemical data resources integrated in the Open PHACTS Discovery Platform.

In this thesis, we developed the Collector Application which will be described in section 3.1.1.

Another way of exploiting the data in the OPS platform is using workflow tools such as KNIME [53] or Pipeline Pilot [54]. The project contributed several KNIME nodes[15] that facilitate the Open PHACTS API access from a KNIME workflow platform. For Pipeline Pilot a set of components have been developed analogously[16]

Recently, some workflows, making use of these components, designed to answer scientifically relevant questions have been published[17] [45].

[15] https://github.com/openphacts/OPS-Knime

[16] https://www.sciencecloud.com/about/open-phacts-component-collection-1-4-released/index.html

[17] http://www.myexperiment.org/groups/1125/content

1.4 Extraction from private preclinical toxicity data

1.4.1 Objectives of eTOX

The eTOX project [23] proposes a strategy to exploit preclinical toxicity data generated for compounds of pharmaceutical interests, extracted from previously unpublished, legacy reports from thirteen European operation-based pharma companies to build predictive toxicity models. The project specifically aims to:

- Share high quality proprietary toxicity data: The data will come from studies conducted under Good Laboratory Practices (GLP). These data from the EFPIA companies will be stored in a database. Data harmonization will be conducted to store these data.

- Development of predictive computational models representing key components of the mechanism leading to a toxic effect. These will be integrated in an overall decision-making tool allowing prediction of *in vivo* toxicity

1.4.2 *In vivo* data Extraction Strategy

Despite since 1981 all preclinical toxicity studied adhere to GLP principles, and high-quality reports are generated, this information is not stored in a structured format in every company. This lack of structured storage makes impossible or very costly to extract even the simplest statistics across different reports in a company. This issue will be mitigated for new studies, after the adoption of the Standard for Exchange of Nonclinical Data (SEND) format[18]. All Raw data of toxicology animal studies started after December 18, 2016 generated to support submission of new drugs to the US Food and Drug Administration must be submitted to the agency using SEND.

Unfortunately, this will not solve the problem for legacy reports and it would be very useful to analyze this data (in company o inter-company) to extract information that can help to learn how to avoid costly failures in the future.

For the inter-company data integration, beyond structuring the data, another important issues arise: chemical structure normalization, terminology normalization, etc. The standard for data collection for the preclinical toxicity reports does not contain any specification of how to structure or normalize the information.

The conclusion is that, for example, none of the 13 companies involved in eTOX project currently has the ability to answer simple questions from their own data such as:

[18] https://www.cdisc.org/standards/foundational/send

- “What type of compound-induced liver toxicity is the most commonly observed in rat across all studies?”
- “What is the translatability of toxicity findings across species?”

Another important issue is that these reports have been generated for regulatory purposes; the goal of this reports was to produce enough information to help regulators to decide about the potential toxicity of drugs in development and decide if they can be tested in humans. So, for example, the data is structured around reports and compounds, but to use the data for toxicity prediction, the viewpoint on the data has to pivot to a cross- compound global view of the data.

Such questions could be answered, in principle, if the data from study reports are put in a structured database. Moreover, more complex questions could be answered.

So one of the primary goals of the eTOX project is to build such inter-company structured database.

The global eTOX strategy can be summarised in the following diagram:

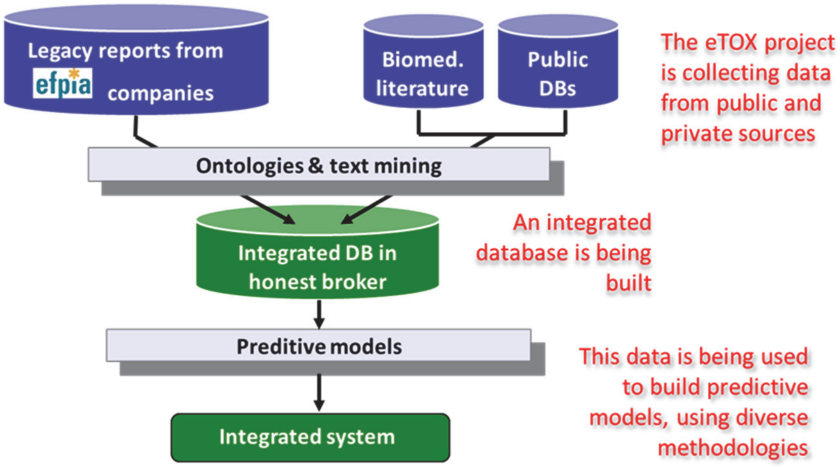

Legacy reports from efpia companies
Biomed. literature
Public DBs
The eTOX project is collecting data from public and private sources
Ontologies & text mining
Integrated DB in honest broker
An integrated database is being built
Preditive models
This data is being used to build predictive models, using diverse methodologies
Integrated system

Figure 8: eTOX strategy

As we can see in the diagram, the main parts of the process are:

- Ontologies & text mining: 3rd party companies have the task of processing the legacy reports and store them in a structured format. During the project, ontologies developed in a collaborative manner allowed to annotate the data in a normalised an precise way.

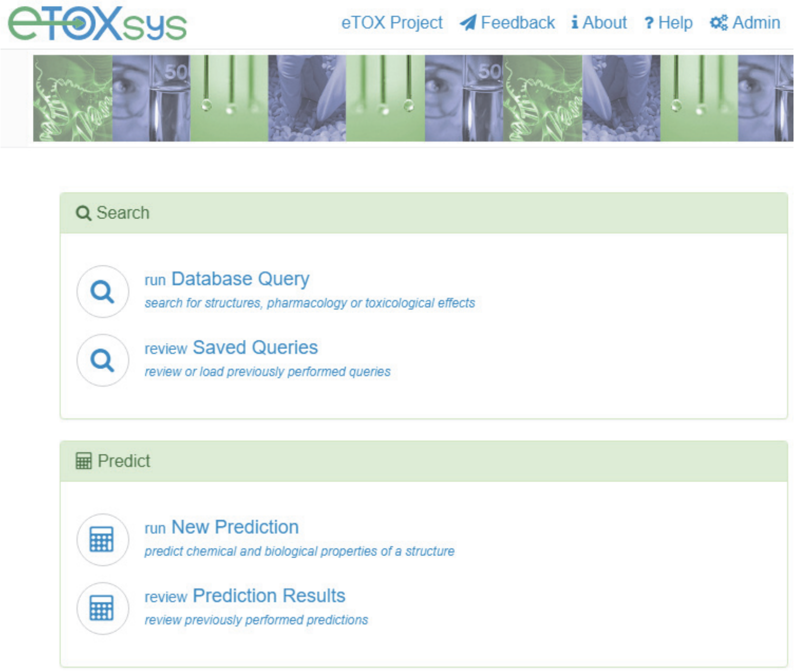
- Integrated DB building: starting from the normalised and structured data, an honest broker built an integrated database that can be accessed by all partners in the project.

- Predictive models: on top of this integrated database and other sources an integrated system (eTOXsys platform)[19] has been built. It implements sophisticated data querying tools, adapted to the needs of the industries participating in the project and several predictive toxicity model.

[19] https://www.etoxsys.com/ https://etoxsys.eu

1.4.3 Results of the project

The project developed the eTOXsys platform, giving access in a single, integrated interface to the database query and to the predictive models. The main interface is shown below:



eTOXsys

eTOX Project Feedback About ? Help Admin

Search

run Database Query
search for structures, pharmacology or toxicological effects

review Saved Queries
review or load previously performed queries

Predict

run New Prediction
predict chemical and biological properties of a structure

review Prediction Results
review previously performed predictions

Figure 9. eTOXsys main screen

The following figure shows the web interface to query the database:

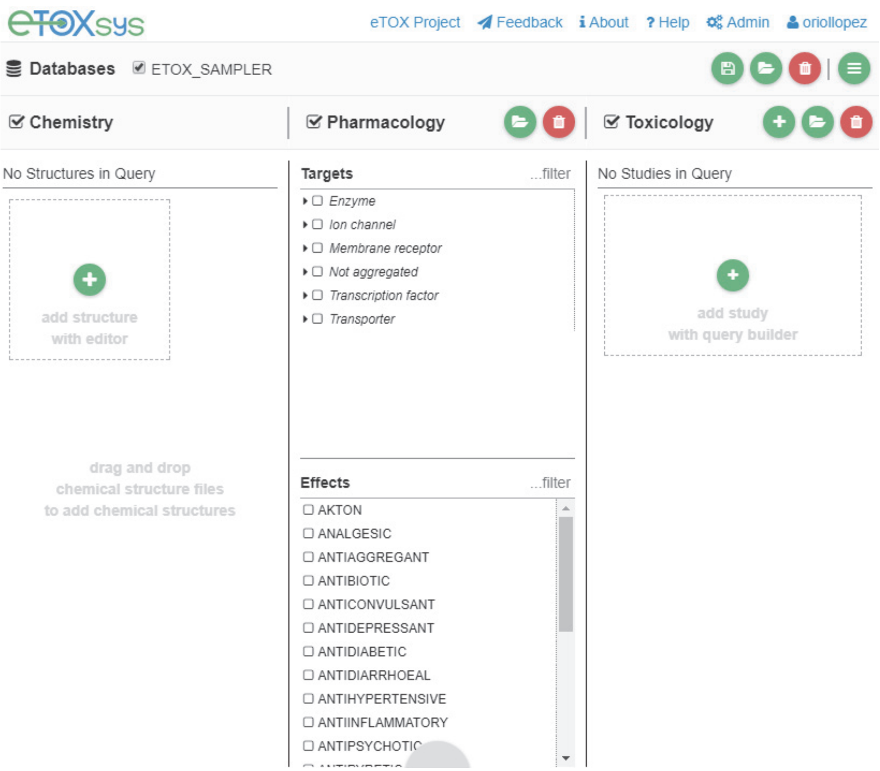
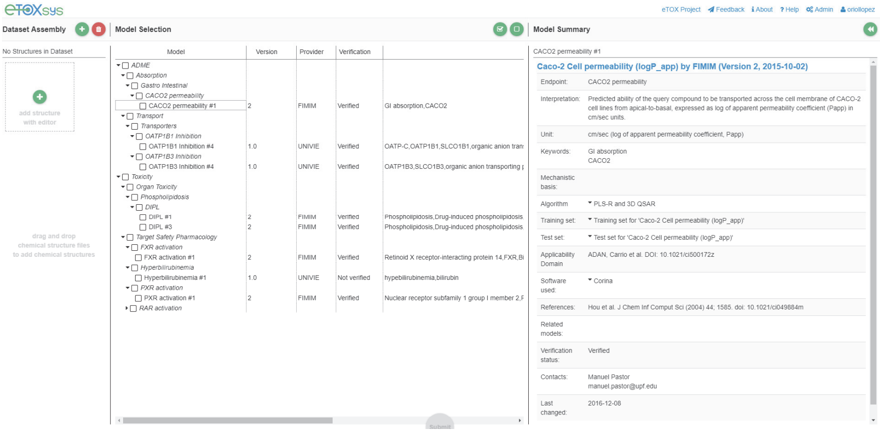


Figure 10: eTOXsys models GUI

The following figure shows the web interface to model prediction module:



Model	Version	Provider	Verification	
☐ *ADME*				
☐ *Absorption*				
☐ *Gastro Intestinal*				
☐ *CACO2 permeability*				
☐ CACO2 permeability #1	2	FIMIM	Verified	GI absorption,CACO2
☐ *Transport*				
☐ *Transporters*				
☐ *OATP1B1 Inhibition*				
☐ OATP1B1 Inhibition #4	1.0	UNIVIE	Verified	OATP-C,OATP1B1,SLCO1B1,organic anion tran:
☐ *OATP1B3 Inhibition*				
☐ OATP1B3 Inhibition #4	1.0	UNIVIE	Verified	OATP1B3,SLCO1B3,organic anion transporting ן
☐ *Toxicity*				
☐ *Organ Toxicity*				
☐ *Phospholipidosis*				
☐ *DIPL*				
☐ DIPL #1	2	FIMIM	Verified	Phospholipidosis,Drug-induced phospholipidosis,
☐ DIPL #3	2	FIMIM	Verified	Phospholipidosis,Drug-induced phospholipidosis,
☐ *Target Safety Pharmacology*				
☐ *FXR activation*				
☐ FXR activation #1	2	FIMIM	Verified	Retinoid X receptor-interacting protein 14,FXR,Bi
☐ *Hyperbilirubinemia*				
☐ Hyperbilirubinemia #1	1.0	UNIVIE	Not verified	hypebilirubinemia,bilirubin
☐ *PXR activation*				
☐ PXR activation #1	2	FIMIM	Verified	Nuclear receptor subfamily 1 group I member 2,F
☐ *RAR activation*				

CACO2 permeability #1

Caco-2 Cell permeability (logP_app) by FIMIM (Version 2, 2015-10-02)

Endpoint:	CACO2 permeability
Interpretation:	Predicted ability of the query compound to be transported across the cell membrane of CACO-2 cell lines from apical-to-basal, expressed as log of apparent permeability coefficient (Papp) in cm/sec units.
Unit:	cm/sec (log of apparent permeability coefficient, Papp)
Keywords:	GI absorption CACO2
Mechanistic basis:	
Algorithm	PLS-R and 3D QSAR
Training set:	Training set for 'Caco-2 Cell permeability (logP_app)'
Test set:	Test set for 'Caco-2 Cell permeability (logP_app)'
Applicability Domain	ADAN, Carrio et al. DOI: 10.1021/ci500172z
Software used:	Corina
References:	Hou et al. J Chem Inf Comput Sci (2004) 44; 1585. doi: 10.1021/ci049884m
Related models:	
Verification status:	Verified
Contacts:	Manuel Pastor manuel.pastor@upf.edu
Last changed:	2016-12-08

Figure 11. eTOXys model prediction GUI

The latest development version includes 202 models, covering a wide range of toxicity endpoints:

Endpoint type	Number of models
Target safety pharmacology	97
Organ toxicity	41
ADME	32
Transporters	27
Carcinogenicity	2
Genotoxicity	2
Phys. chem. properties	1

eTOX [55] based its modelling strategy on the previously defined concept of AOP (See 1.1.4). AOPs define a Molecular Initiator Events (MIE) that produce a series of linked causally events that finally produce an adverse effect. It is important to note that the chain of effects derived are independent of the toxicant since the toxicant only triggers the MIE (or MIEs). The main point in such framework is to determine if the compound triggers the MIE. So for this reason we can use QSAR methods to predict the risk of triggering the MIE.

For modelling complex endpoints such as hepatoxicity, several mechanisms described by his own AOP triggered by different MIEs have to be taken into account. The eTOX approach was to predict single MIEs (first order models) and then combine them to predict the in vivo outcome (second order models) [55]:

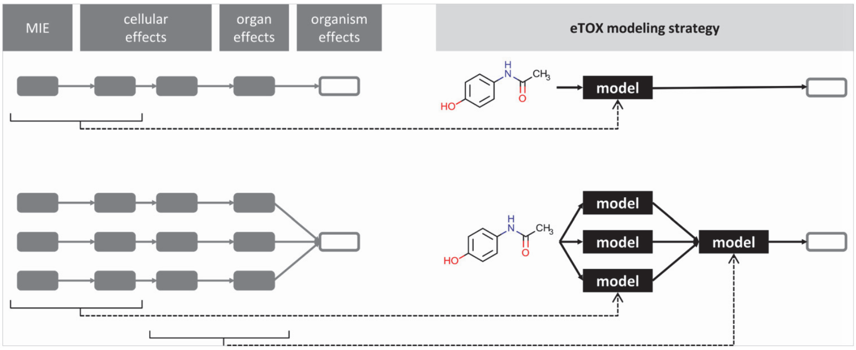


Figure 12: eTOX modelling strategy [56]

One example of such top level model is the QT prolongation predictive model that relies on two first order models to predict blockade of two ion channels [56].

Another result of the project is the *in vivo* preclinical database described above. It contains, in the latest version, 1947 compounds and 8047 associated repeat dose studies[20]. This data has been derived from proprietary data donated by the 13 pharmaceutical eTOX partners.

The database structure is shown below:

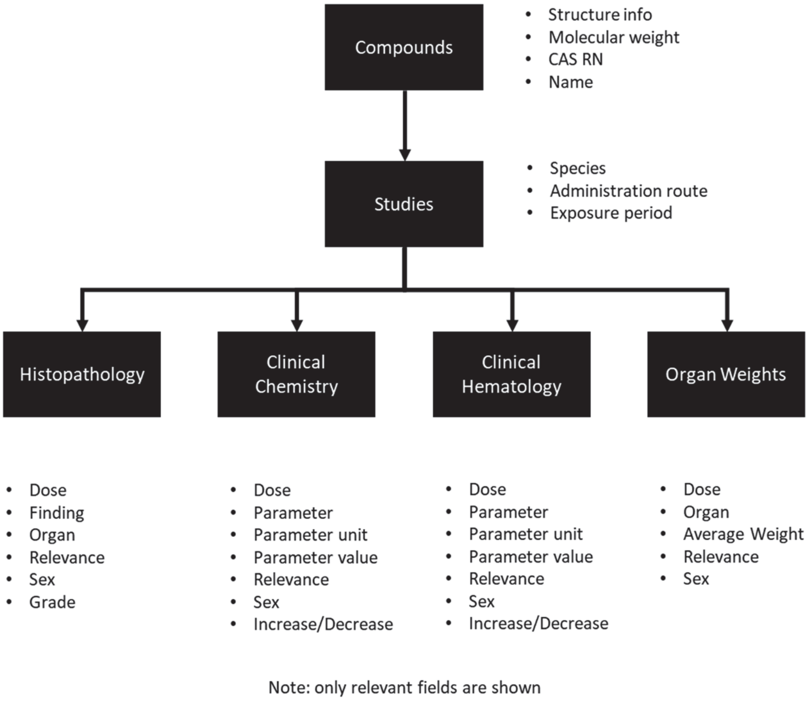


Figure 13: eTOX in vivo database

[20] 2016.3 release

The project entered the sustainability phase at the beginning of 2017, and the results are accessible only to the eTOX partners, or through a private commercial agreement with the eTOX board.

1.4.4 Database construction

The original strategy was to store all the reports information in Vitic, a database developed by Lhasa. Lhasa developed the database for eTOX using VITIC platform. Lhasa acted as an honest broker to store all the chemical and preclinical provided by the EFPIA partners involved. They provided EFPIA partners the tools forsending and integrating their preclinical data.

At the end of the eTOX project the data was made accessible through an integrated web application (eTOXsys) that allows querying the data according to the specific needs of the pharmaceutical companies involved in the project. For more details on the schema and the volume of the database see 3.1.2

This database allows accessing the original information of the preclinical studies in a integrated and structured way. For every compound and report the database stores thousands of data points with the measurements of the effects of the animal experiments with all the relevant attributes (species, administration route, duration, relevance of the effect, quantitative/qualitative observation).

1.4.5 Ontologies and Normalization

Extracting preclinical toxicity data from toxicity reports was a major challenge. These reports have been compiled by different companies over the years [23], so there is no homogeneous way to annotate such toxicity data.

Specifically, during the project, an important task was to normalise the terminology used in the reports (for example species, administration route etc.) and to develop ontologies to annotate with precision the histopathological data (organ ontology and histopathological ontology[21]).

These ontologies have been developed collaboratively by the experts in the project using OntoBrowser [57].

1.4.6 From data to knowledge

Despite building a high-quality relational database (even storing ontology annotated data), the data present there in its original form cannot be used for toxicity modelling. Modelling needs to work on comparable information, expressing properties of the compounds. In contrast, the database

[21] Released under Apache License, Version 2.0 https://github.com/Novartis/hpath

contains observations made at different dose, time, administration routes and on different species.

We need data series of information for top-level toxicity endpoints summarised at compound level, describing the detailed, fine grained, information of the findings in the database.

2 OBJECTIVES

The general objective of the project is:

- Development of computational tools for the automatic extraction, curation and normalization of training series from open sources of biomedical data as well as from corporate databases, in formats that allow their direct use for the development of predictive *in silico* models

The specific objectives are:

- Development of data extraction tools connected to a selected collection of biomedical data sources. (Collector)

- Development of data extraction tools connected to relevant private databases used in drug development. Specifically, we will develop tools to extract information from repeated-dose toxicity studies reports compiled in the framework of the eTOX project.

3 RESULTS

3.1 Overview of results

3.1.1 EXTRACTION OF BIOACTIVITY DATA FROM PUBLIC DATA SOURCES FOR MODELLING PURPOSES

3.1.1.1 THE NEED. COLLECTOR PROPOSAL

We have developed a tool, Collector [24] to gather data from public data sources with the aim of building QSAR-like predictive models. Specifically, Collector was developed to build predictive models for toxicity endpoints, in the framework of the eTOX project, for which we need obtaining pharmacological/toxicological data (IC_{50}, pK_D, etc.) for the largest possible series of compounds. The process involves data gathering from public data sources as well as a process of data curation to discard data according to quality/relevance criteria.

According to this need, the developed tool:

- Uses the Open PHACTS platform to extract relevant data to build toxicity models in eTOX.
- Includes data curation features that will filter the data extracted from Open PHACTS
- Includes a web-based GUI to create and monitor data extraction and data curation jobs.

- Includes a web-based GUI to browse the results obtained.

3.1.1.2 RESULTS OBTAINED

- We developed Collector. The source code is available as open source software, distributed under GPLv3 license. See Section 3.2.2.8. It can be easily installed on any modern Linux 64 distribution.

- Collector relies on the framework of linked data and semantic web and, specifically, on the Open PHACTS platform. For this reason, Collector will benefit from any new datasets included in the platform.

- The data obtained by Collector is linked data so it benefits from the principles of unique identified concepts such as targets, compounds, etc via URIs. See Section 4 for discussion and limitations

- The linked data obtained allows to navigate through links to other referenced data sets: from target to genes, from genes to pathways, from compounds to related compounds, etc. See Section 4 for discussion and limitations

- Collector provides both a GUI and a command line interface for defining and executing such extractions in an easy way. See Sections 3.2.2.5 and 3.2.2.6

- A demo instance of Collector is hosted on-line at http:///collector.upf.edu, where it is available for the research community

3.1.1.3 eTOX MODELS

We used Collector to extract training series used for developing different toxicity models in the eTOX project. See the supplementary material of the article in Section 3.2 for more details of the models developed.

We used an in-house developed platform (eTOXlab) [26] to implement and deploy such models in eTOXsys[22]. See Sections 3.4 and 3.4.2 for more details.

3.1.1.4 eTOX LQT MODELS

We deployed and implemented some of the LQT models developed in our group in a web server accessible at the following address:

http://etoxlab-lqt.upf.edu/

[22] https://www.etoxsys.com/

3.1.2 EXPLOITING PRECLINICAL DATA FOR TOXICITY MODELLING PURPOSES

3.1.2.1 PRINCIPLES AND STRATEGY

We conceived, designed, implemented and tested a new methodology to extract data from RDT reports to obtain data series for toxicity modelling purposes.

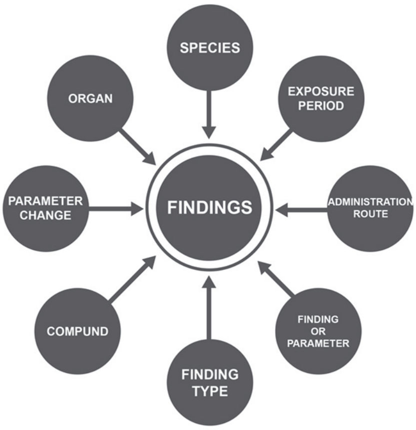
The approach has been developed and tested in the framework of the eTOX project [23]. See Section 1.4 for the background of the project.

As we mentioned before, despite the effort to extract the data from preclinical RDT reports the data is not amenable to model *in vivo* toxicity, since the structure of the database does not allow to extract comparable toxicity annotations for series of compounds.

Therefore, we decided to design, develop and implement a methodology to provide such data. It is based on the following principles:

- Some of the data (histopathological data) in the preclinical reports was normalized using two ontologies: the organ ontology (to model organs and his parts) and pathology ontology (to model findings in organs and tissues). We developed a mechanism to infer new derived data based on these ontologies.

- The data from preclinical reports is multidimensional (specie, exposure period, administration route, etc) and fine grained (it contains individual measurements of effects). It stores millions of data points with multiple dimensions. Schematically:



  Specifically, a measured quantitative or qualitative fact is annotated in the different dimensions: species, organs,…

Following this OLAP approach, to obtain relevant data for modelling we decided to process this data using the principles and methods of OLAP systems [58].

To obtain data for modelling at compound level using QSAR methods (see 1.2.3 for background), the analysis must be extracted from studies carried out in similar conditions, and we need to extract a subcube of the multidimensional data (with certain homogeneity) and aggregate this data at compound level. There is a trade-off between the homogeneity of the studies and the number of compounds extracted: more homogeneous extractions yield smaller datasets and vice versa.

- We defined a method to compute toxicity scorings at compound level for *in vivo* endpoints. It was applied for generating both quantitative and quantitative scorings.

The methodology was described in more detail in Section 3.3.

3.1.2.2 Example Hepatotoxicity

This methodology was used to model *in vivo* Hepatotoxicity in a collaboration with SANOFI. Specifically, we extracted the data using our method to obtain compound-level data for modelling 3 top-level endopints: 'non-neoplasic proliferative lesions' (PRO), 'inflammatory liver changes' (INF) and 'degenerative lesions' (DEG).

We computed two kinds of scoring at compound level: qualitative (using a logical OR on the observed findings) and qualitative (using the minimal dose of the observed findings, transformed into minus logarithmic scale) to generate “pseudo-LOEL”.

We compared our quantitative scoring (pseudo-LOEL) values obtained from liver-related findings with NOAEL values reported in the eTOX database for some compounds, and they exhibit the expected correlation, further justifying the relevance and usefulness of the proposed quantitative scoring.

We developed several models based on the qualitative scoring, the quality of the models is not excellent, but is acceptable if we consider the complexity of the *in vivo* endpoints based on many different mechanism of liver toxicity that are being described and the simplicity of the model approach described here.

We also developed models based on the quantitative scorings with slightly worse results.

3.2 Article “An automated tool for obtaining QSAR-ready series of compounds using semantic web technologies”

3.2.1 Article

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Databases and ontologies

Databases and ontologies

An automated tool for obtaining QSAR-ready series of compounds using semantic web technologies

Oriol López-Massaguer, Ferran Sanz, Manuel Pastor*

Research Programme on Biomedical Informatics (GRIB), Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Dept. of Experimental and Health Sciences, Universitat Pompeu Fabra, C/. Dr. Aiguader 88, 08003 Barcelona, Spain.

*To whom correspondence should be addressed.

Associate Editor: Dr. John Hancock

Collector [59] is an application for obtaining series of compounds annotated with bioactivity data, ready to be used for the development of quantitative structure-activity relationships (QSAR) models. The tool extracts data from the ‘Open Pharmacological Space’ (OPS) developed by the Open PHACTS project, using as input a valid name of the biological target. Collector uses the OPS ontologies for expanding the query using all known target synonyms and extracts compounds with bioactivity data against the target from multiple sources. The extracted data can be filtered to retain only drug-like compounds and the bioactivities can be automatically summarized to assign a single value per compound, yielding data ready to be used for QSAR modeling. The data obtained is locally stored facilitating the traceability and auditability of the process. Collector was used successfully for the development of models for toxicity endpoints within the eTOX project.

López-Massaguer O, Sanz F, Pastor M. An automated tool for obtaining QSAR-ready series of compounds using semantic web technologies. Bioinformatics. 2018 Jan 1;34(1):131-133. doi: 10.1093/bioinformatics/btx566

3.3 Article “Generating modelling data from repeat-dose toxicity reports”

3.3.1 Article

Generating modelling data from repeat-dose toxicity reports

Submitted to Toxicological Sciences on 4th of August 2017.

Accepted with minor revisions 25th September 2017.

Oriol López-Massaguer[*], Kevin Pinto-Gil[†], Ferran Sanz[‡], Alexander Amberg[§], Lennart T. Anger[¶], Manuela Stolte[||], Carlo Ravagli[|||], Philippe Marc[||||] and Manuel Pastor [#,1]

[*] Research Programme on Biomedical Informatics (GRIB), Institut Hospital del Mar d’Investigacions Mèdiques (IMIM), Dept. of Experimental and Health Sciences, Universitat Pompeu Fabra, C/. Dr. Aiguader 88, 08003 Barcelona, Spain. E-mail: oriol.lopez@upf.edu.

[†] Research Programme on Biomedical Informatics (GRIB), Institut Hospital del Mar d’Investigacions Mèdiques (IMIM), Dept. of Experimental and Health Sciences, Universitat Pompeu Fabra, C/. Dr. Aiguader 88, 08003 Barcelona, Spain. E-mail: kevin.pinto@upf.edu.

[‡] Research Programme on Biomedical Informatics (GRIB), Institut Hospital del Mar d’Investigacions Mèdiques (IMIM), Dept. of Experimental and Health Sciences, Universitat Pompeu Fabra, C/. Dr. Aiguader 88, 08003 Barcelona, Spain. E-mail: ferran.sanz@upf.edu.

[§]Sanofi-Aventis Deutschland GmbH, R&D, Preclinical Safety, 65926 Frankfurt am Main, Germany. E-mail: Alexander.Amberg@sanofi.com.

[¶]Sanofi-Aventis Deutschland GmbH, R&D, Preclinical Safety, 65926 Frankfurt am Main, Germany. E-mail: Lennart.Anger@sanofi.com

[||]Sanofi-Aventis Deutschland GmbH, R&D, Preclinical Safety, 65926 Frankfurt am Main, Germany. E-mail: Manuela.Stolte@sanofi.com

[lll] Translational Medicine, Novartis Institute for Biomedical Research, CH-4002 Basel, Switzerland. E-mail: carlo.ravagli@novartis.com.

[llll] Translational Medicine, Novartis Institute for Biomedical Research, CH-4002 Basel, Switzerland. E-mail: philippe.marc@novartis.com.

[#] Research Programme on Biomedical Informatics (GRIB), Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Dept. of Experimental and Health Sciences, Universitat Pompeu Fabra, C/. Dr. Aiguader 88, 08003 Barcelona, Spain. E-mail: manuel.pastor@upf.edu.

1 Corresponding author

López-Massaguer O, Pinto-Gil K, Sanz F, Amberg A, Anger LT, Stolte M, Ravagli C, Marc P, Pastor M. Generating Modeling Data From Repeat-Dose Toxicity Reports. Toxicol Sci. 2018 Mar 1;162(1):287-300. doi: 10.1093/toxsci/kfx254

3.3.2 Additional details

3.3.2.1 Method summary

The method proposed can be summarized in the following steps:

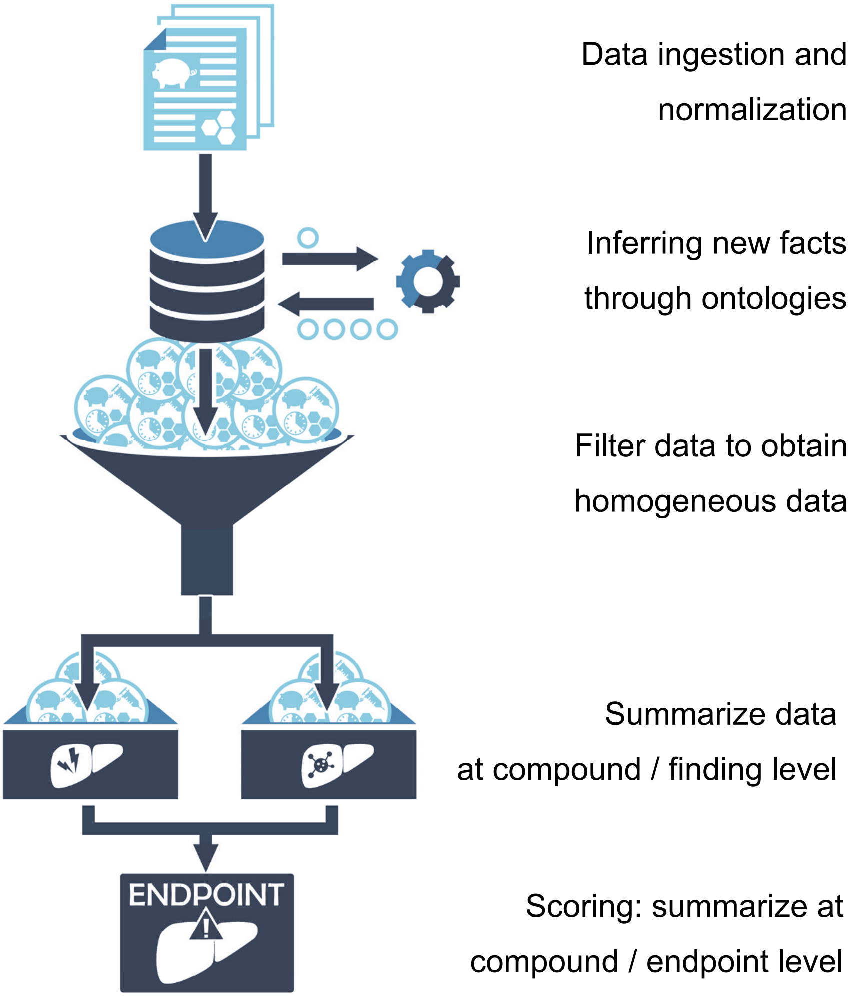

Data ingestion and normalization
Inferring new facts through ontologies
Filter data to obtain homogeneous data
Summarize data at compound / finding level
ENDPOINT
Scoring: summarize at compound / endpoint level

Figure 17: *In vivo* data extraction method summary

The method starts from the database containing preclinical data from eTOX project described in Section 1.4.4

But we need to extract from this high volume of complex data relevant information that can be used for modelling purposes.

3.3.2.2 FINDING DEFINITION

The starting point is to define what we understand by “finding” in the context of preclinical data.

A finding is an observation corresponding to an effect of the treatment applied (in this case drug testing).

Specifically, we consider a finding any treatment related (based on expert judgment found in the preclinical testing reports) alteration of a histopathological finding or an alteration (increase or decrease) in a quantitative parameter measurement (clinical chemical finding, etc.).

3.3.2.3 INFERENCE

Histopathological data stored in the data base is annotated with two ontologies. Any histopathological data point is annotated with the finding (ex: Necrosis) and the organ where it was found (ex: Liver).

Ontologies used in this project were developed with semantic web standards [39,44,68].

A critical distinction in description logics (the formalism underlying semantic web standards) is the distinction between:

- TBox (Terminological Box): It formalizes the terminological knowledge of the domain (ontologies). It is purely intensional knowledge.
- ABox (Assertion Box): It formalizes the factual knowledge. Specifically we can assert that an individual belongs to a concept or relation assertions. It is extensional knowledge

In our case, the Tbox is constituted by the two ontologies: histopathological ontology and anatomical ontology.

On the other hand we interpret the facts stored in the relational database storing preclinical data as our ABox.

This interpretation is not strange since relational database have a classical model based on first order predicate calculus [69] and in fact the query languages of relational database can be interpreted as a restricted form of inference.

3.3.2.3.1 Ontology based query reformulation

Our goal is to derive new facts from the stored data by using the ontologies used to annotate the data. It can be done following several approaches:

- We can use TBox interference to derive new facts at ontology level using standard OWL reasoners. However, this reasoning only covers TBox inference but in our case, we need ABox inference and this approach is complex and probably not scalable.

- Another method could be to transform the whole relational database to triples and store it in a triplestore. Apart from the fact that performing this conversion, SPARQL triple stores are not as mature as relational databases in performance terms, especially if we want to apply ABox inference.

- We implemented the inference using SQL recursive queries. This has major advantages for us:

    - Integrates the inference mechanism in database query language. No need of a separated inference engine from que query engine.

    - As the inference is integrated with the query mechanism, we do not need to pre-compute inferred facts and store them. We only infer the data to respond the query of interest,

    - Our need is to have a flexible and general mechanism of filtering to exploit the multidimensional nature of the data. See 3.1.2.1.

3.3.2.4 Filtering: Data homogeneity

One critical point of obtaining data for modelling is to get datasets that have a balance between size and homogeneity.

If we filter data to a narrow condition (only one species, or administration route) the data set will be small for modelling, on the contrary, if we widen the filtering conditions the data heterogeneity will affect the modelling due to the data variability.

We implemented a flexible and general filtering mechanism to our multidimensional data by applying the classical analytical query operators for OLAP data cubes [69] like: slice, dice, roll-up, drill-down.

However, due to the relatively modest size of our database (about 10 million rows) we do not need to use specialised OLAP database servers but we used PostgreSQL database server directly.

The flexible query mechanism allowed us to perform different data extractions for various endpoints and refine interactively our data extractions until the datasets obtained fulfil our needs.

3.3.2.5 Toxicological Profiles. Scoring

After data is obtained by applying the inference and filtering mechanisms described, the findings obtained must be further aggregated to higher level toxicity endpoints suitable for modelling. We applied this strategy in our example of Hepatotoxicity modelling. See Section 3.1.2.2.

For every compound and toxicological profile we computed two kinds of scorings: qualitative (if any of the findings defining the profile exists (logical OR)) and quantitative (minimal dose at which we found any of the effects corresponding to the profile).

We finally obtain a data series of compounds annotated with the different toxicological endpoints (quantitative or qualitative) suitable for modelling.

3.4 Other Results

3.4.1 eTOXlab

Carrió *et al. Journal of Cheminformatics* (2015) 7:8
DOI 10.1186/s13321-015-0058-6

Journal of Cheminformatics

SOFTWARE — Open Access

eTOXlab, an open source modeling framework for implementing predictive models in production environments

Pau Carrió, Oriol López, Ferran Sanz and Manuel Pastor*

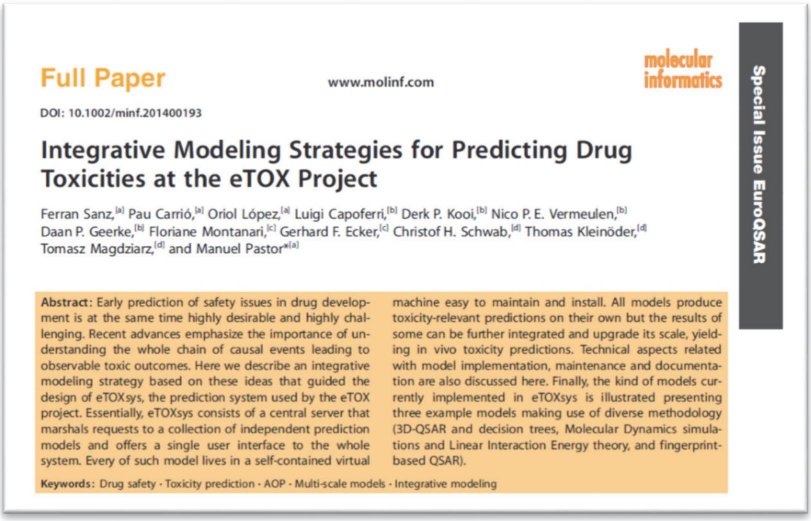
eTOXlab[26] is an open source modeling framework designed to be used at the core of a self-contained virtual machine that can be easily deployed in production environments, providing predictions as web services. eTOXlab consists on a collection of object-oriented Python modules with methods mapping common tasks of standard modeling workflows. This framework allows building and validating QSAR models as well as predicting the properties of new compounds using either a command line interface or a graphic user interface (GUI).

We developed eTOXlab in our group and it is our standard tool to develop and deploy QSAR models. We applied it along this thesis: to implement the QSAR models for hepatotoxicity based on preclinical data described in Section 3.3.1, to implement and deploy the LQT models developed in the framework of the eTOX project (See Section 3.1.1.4) and the models developed in eTOX using Collector to obtain data (See Section 3.1.1.3).

Oriol López Massaguer contributions:

- Contribute to the system Overall design of eTOXlab platform.
    - Homepage http://phi.imim.es/envoy/
    - Source code in https://github.com/phi-grib/eTOXlab
- Setup of the virtual machine which contains the implementation of eTOXlab
    - Homepage http://phi.imim.es/envoy/
- Implementation of Padel [70] descriptors computation subsystem
    - Source code in https://github.com/phi-grib/PaDEL-descriptor-ws
- Implementation of prediction subsystem API
    - Source code in https://github.com/phi-grib/eTOXlabWS

3.4.2 INTEGRATIVE MODELLING IN THE ETOX PROJECT



Full Paper

www.molinf.com

molecular informatics

Special Issue EuroQSAR

DOI: 10.1002/minf.201400193

Integrative Modeling Strategies for Predicting Drug Toxicities at the eTOX Project

Ferran Sanz,[a] Pau Carrió,[a] Oriol López,[a] Luigi Capoferri,[b] Derk P. Kooi,[b] Nico P. E. Vermeulen,[b] Daan P. Geerke,[b] Floriane Montanari,[c] Gerhard F. Ecker,[c] Christof H. Schwab,[d] Thomas Kleinöder,[d] Tomasz Magdziarz,[d] and Manuel Pastor*[a]

Abstract: Early prediction of safety issues in drug development is at the same time highly desirable and highly challenging. Recent advances emphasize the importance of understanding the whole chain of causal events leading to observable toxic outcomes. Here we describe an integrative modeling strategy based on these ideas that guided the design of eTOXsys, the prediction system used by the eTOX project. Essentially, eTOXsys consists of a central server that marshals requests to a collection of independent prediction models and offers a single user interface to the whole system. Every of such model lives in a self-contained virtual machine easy to maintain and install. All models produce toxicity-relevant predictions on their own but the results of some can be further integrated and upgrade its scale, yielding in vivo toxicity predictions. Technical aspects related with model implementation, maintenance and documentation are also discussed here. Finally, the kind of models currently implemented in eTOXsys is illustrated presenting three example models making use of diverse methodology (3D-QSAR and decision trees, Molecular Dynamics simulations and Linear Interaction Energy theory, and fingerprint-based QSAR).

Keywords: Drug safety · Toxicity prediction · AOP · Multi-scale models · Integrative modeling

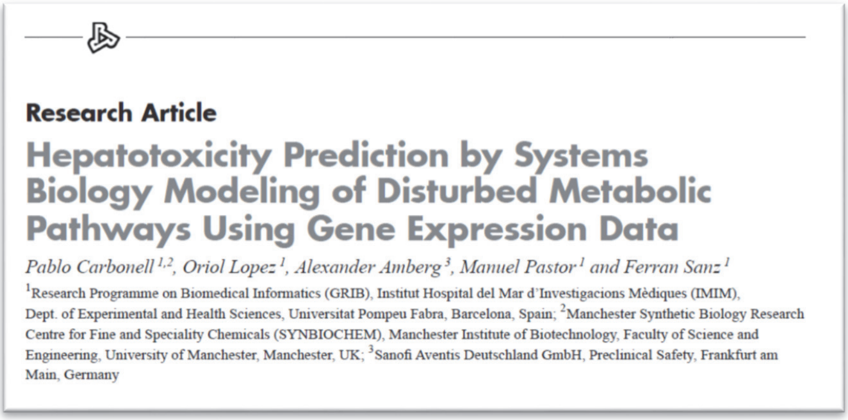
Integrative Modelling in the eTOX project [55] . Early prediction of safety issues in drug development is at the same time highly desirable and highly challenging. Recent advances emphasize the importance of understanding the whole chain of causal events leading to observable toxic outcomes. Here we describe an integrative modeling strategy based on these ideas that guided the design of eTOXsys, the prediction system used by the eTOX project. Essentially, eTOXsys consists of a central server that marshals requests to a collection of independent prediction models and offers a single user interface to the whole system. Every of such model lives in a self-contained virtual machine easy to maintain and install. All models produce toxicity-relevant predictions on their own but the results of some can be further integrated and upgrade its scale, yielding *in vivo* toxicity.

The eTOX modelling strategy and architecture is the foundation for deploying the toxicity models developed during the project. It includes also the models described in this thesis in Sections 3.1.1.3 and 3.1.1.4

Oriol López Massaguer contributions:

- Documentation of models in eTOXvault database and eTOXvault WS in eTOXsys:
    - eTOXvault ws source code: https://github.com/phi-grib/eTOX-vault-ws
    - Data model of eTOXvault
- Models LQT VM machine Web Service setup:
    - Server: http://etoxlab-lqt.upf.edu/
    - Code: https://github.com/phi-grib/eTOXlabWS
    - LQT model is based on the group previous publication [56]

3.4.3 HEPATOXICITY PREDICTION BY SYSTEMS BIOLOGY APPROACH



Research Article

Hepatotoxicity Prediction by Systems Biology Modeling of Disturbed Metabolic Pathways Using Gene Expression Data

Pablo Carbonell[1,2], Oriol Lopez[1], Alexander Amberg[3], Manuel Pastor[1] and Ferran Sanz[1]

[1]Research Programme on Biomedical Informatics (GRIB), Institut Hospital del Mar d'Investigacions Mèdiques (IMIM), Dept. of Experimental and Health Sciences, Universitat Pompeu Fabra, Barcelona, Spain; [2]Manchester Synthetic Biology Research Centre for Fine and Speciality Chemicals (SYNBIOCHEM), Manchester Institute of Biotechnology, Faculty of Science and Engineering, University of Manchester, Manchester, UK; [3]Sanofi Aventis Deutschland GmbH, Preclinical Safety, Frankfurt am Main, Germany

The study [71] applies a systems biology approach for the *in silico* predictive modeling of drug toxicity on the basis of high-quality preclinical drug toxicity data with the aim of increasing the mechanistic understanding of toxic effects of compounds at different levels (pathway, cell, tissue, organ). The model development was carried out using 77 compounds for which gene expression data for treated primary human hepatocytes is available in the LINCS database and for which rodent *in vivo* hepatotoxicity information is available in the eTOX database. The data from LINCS were used to determine the type and number of pathways disturbed by each compound and to estimate the extent of disturbance (network perturbation elasticity), and were used to analyze the correspondence with the *in vivo* information from eTOX. Predictive models were developed through this integrative analysis, and their specificity and sensitivity were assessed.

The quality of the predictions was determined on the basis of the area under the curve (AUC) of plots of true positive vs. false positive rates (ROC curves). The ROC AUC reached values of up to 0.9 (out of 1.0) for some hepatotoxicity endpoints. Moreover, the most frequently disturbed metabolic pathways were determined across the studied toxicants. They included, e.g., mitochondrial beta-oxidation of fatty acids and amino acid metabolism. The process was exemplified by successful predictions on various statins. In conclusion, an entirely new approach linking gene expression alterations to the prediction of complex organ toxicity was developed and evaluated.

In this work we validated the systems biology approach used by comparing the results obtained by performing metabolic pathways perturbations in the liver with *in vivo* preclinical data obtained in the framework of the eTOX project.

Oriol López Massaguer contributions:

- Extraction of *in vivo* data from the eTOX database and computation of toxicity scorings.
- Source code data extractions developed
    - https://github.com/phi-grib/Hepatotoxicity-SysBio-DataExtraction (Data extraction)
    - https://github.com/phi-grib/Hepatoxicity-SysBio (SysBio Hepatotoxicity predicion)

4 DISCUSSION

There is an increasing need to improve *in silico* predictive methods to assess drug toxicity. This need comes from both economic reasons and ethical reasons.

The work presented here contributes to predict *in vivo* toxicity. It was illustrated by an study, in collaboration with SANOFI where it was applied to predict *in vivo* hepatotoxicity endpoints.

To our knowledge, the method described here is the first automated procedure to exploit rich *in vivo* preclinical databases, like the one generated by the public-private partnership of the eTOX project. However, we consider this application only a first step as we think that it can be applied in the future to other similar databases and the definition of the toxicity endpoints can be improved.

Our method applies a novel technique to combine information extraction (through SQL) and combining it with the ontology-based reasoning of the annotated data (through SQL recursive queries) and it is entirely general. It can be applied to any relational database (supporting SQL:99) that stores ontology annotated information.

In particular, we think that our approach can be applied in other biomedical databases as in this domain data is frequently annotated with ontologies: Gene Ontology, CHEBI, etc (see [72] for a comprehensive list). For example, Gene Ontology is routinely used to enrich gene information. Another potential use is to apply it to the ChEMBL database which is annotated with several ontologies.

We have large amounts of data annotated with ontologies that have been built with great effort but we lack of general tools to exploit them in combination with data from the ontologies. We know that gene enrichment is usually performed on genomic data but we are not aware of generally applicable strategies for carrying out this ontology data enrichment.

Another important data source for modeling toxicity are public databases but, as we explained in this work, this is not an easy task for various reasons. The FAIR initiative defines a set of principles which might help to overcome this situation and in this framewok the Open PHACT project was developed.

Open PHACTS is a very ambitious project, and despite having integrated several relevant data sources for biomedical research it is not clear that, in the current stage, it is a better alternative than use directly the original data sources.

Our tool Collector was developed in this framework to extract data for *in silico* toxicity modelling. It is an example of a simple an easy tool to obtain QSAR ready data series. But we think the decision of encapsulating the semantic web data stored in the Open PHACTS platform behind a standard web API, and exposing only this access is a serious limitation. If the developer don’t have full access to an SPARQL endpoint most of the advantages of using semantic web technologies are not available. We think for example in complex path queries or reasoning based queries.

Another parallel shortcoming is that SPARQL triple stores are new compared to relational database systems (over 40 years and research and development) and the maturity and performance of triple stores is behind their relational counterparts.

The same fact of long pervasiveness of relational databases makes that most of the databases in our domain (public, proprietary) use relational technology. Therefore, it requires a great effort to convert them to semantic web triple stores format.

In our opinion, a more pragmatic approach would have been to mix relational database technology with ontologies and reasoning as we demonstrated for the extraction of *in vivo* preclinical data.

5 CONCLUSIONS

1. We developed Collector, a tool that compiles series of compounds annotated with bioactivities, ready to obtain QSAR models, using semantic web technologies

2. We have conceived, designed, implemented and tested a new method to extract data from repeated-dose toxicity reports that is usable for building models of in vivo toxicity endpoints.

3. This approach was tested by obtaining quantitative and qualitative toxicity scoring for liver toxicity endpoints.

4. The suitability of these scorings for modeling purposes were confirmed by the development of QSAR models as well as by comparing them with reported NOAELs.

5. We developed a novel method that performs ontology-based inference on top of relational databases. The method combines inference and query in a single step and thus the method can be applied to large datasets.

6 REFERENCES

1 DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&D costs. J Health Econ 2016; 47: 20–33 Available from: http://www.ncbi.nlm.nih.gov/pubmed/26928437

2 PHRMA. Industry profile. 2016 Available from: http://phrma-docs.phrma.org/sites/default/files/pdf/biopharmaceutical-industry-profile.pdf

3 Morrison JS, Mudra DR, Editors DMB. Translating Molecules into Medicines. Cham: Springer International Publishing, 2017 Available from: http://link.springer.com/10.1007/978-3-319-50042-3

4 Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Publ Gr 2012; 11 Available from: https://www.ncbi.nlm.nih.gov/pubmed/22378269

5 Schulze U, Baedeker M, Chen YT, Greber D. R&D productivity: on the comeback trail. Nat Rev Drug Discov 2014; 13: 331–332 Available from: http://dx.doi.org/10.1038/nrd4320

6 Waring MJ, Arrowsmith J, Leach AR, Leeson PD, Mandrell S, Owen RM, Pairaudeau G, Pennie WD,

Pickett SD, Wang J, Wallace O, Weir A. An analysis of the attrition of drug candidates from four major pharmaceutical companies. 2015; Available from: https://www.nature.com/nrd/journal/v14/n7/pdf/nrd4609.pdf

7 Cases M, Briggs K, Steger-Hartmann T, Pognan F, Marc P, Kleinöder T, Schwab CH, Pastor M, Wichard J, Sanz F. The eTOX data-sharing project to advance in Silico drug-induced toxicity prediction. Int J Mol Sci 2014; 15: 21136–21154 Available from: http://www.mdpi.com/1422-0067/15/11/21136/

8 Keisu M, Andersson TB. Drug-Induced Liver Injury in Humans: The Case of Ximelagatran. In: Handbook of experimental pharmacology. 2010: 407–418 Available from: http://www.ncbi.nlm.nih.gov/pubmed/20020269

9 Jüni P, Nartey L, Reichenbach S, Sterchi R, Dieppe PA, Egger M. Risk of cardiovascular events and rofecoxib: cumulative meta-analysis. Lancet 2004; 364: 2021–2029 Available from: http://linkinghub.elsevier.com/retrieve/pii/S0140673604175144

10 Hussaini SH, Farrington EA. Idiosyncratic drug-induced liver injury: an update on the 2007 overview. Expert Opin Drug Saf 2014; 13: 67–81 Available from:

http://www.tandfonline.com/doi/full/10.1517/14740338.2013.828032

11 Roth RA, Ganey PE. Intrinsic versus Idiosyncratic Drug-Induced Hepatotoxicity—Two Villains or One? J Pharmacol Exp Ther 2010; 332 Available from: http://jpet.aspetjournals.org/content/332/3/692.short

12 Olsson IAS. The 3Rs principle – mind the ethical gap! Altex Proc 2012; Available from: http://www.forskningsdatabasen.dk/en/catalog/2193075412

13 Agencia Europea para la Seguridad y la Salud en el Trabajo. Regulation (EC) No 1907/2006 - REACH - Salud y seguridad en el trabajo - EU-OSHA. Regul No 1907/2006 - Reach 2006; Available from: https://osha.europa.eu/es/legislation/directives/regulation-ec-no-1907-2006-of-the-european-parliament-and-of-the-council

14 European parliameient and European council. Regulation (EC) No 1223/2009. 2009; Available from: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2009:342:0059:0209:en:PDF

15 EPA. Benchmark Dose Technical Guidance - EPA/100/R-12/001. 2012 Available from:

https://www.epa.gov/sites/production/files/2015-01/documents/benchmark_dose_guidance.pdf

16 What Is Point of Departure (POD) and How to Use It to Calculate Toxicological Reference Dose (RfD). Available from: http://www.chemsafetypro.com/Topics/CRA/What_is_Point_of_Departure_(POD)_in_Toxicology_and_How_to_Use_It_to_Calculate_Reference_Dose_RfD.html

17 Council NR. Toxicity Testing in the 21st Century. Washington, D.C.: National Academies Press, 2007 Available from: http://www.nap.edu/catalog/11970

18 Bracken MB. Why animal studies are often poor predictors of human reactions to exposure. J R Soc Med 2009; 102: 120–122 Available from: http://www.ncbi.nlm.nih.gov/pubmed/19297654

19 Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, Mount DR, Nichols JW, Russom CL, Schmieder PK, Serrrano JA, Tietge JE, Villeneuve DL. Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. Environ Toxicol Chem 2010; 29: 730–741 Available from: http://doi.wiley.com/10.1002/etc.34

20 Vinken M, Knapen D, Vergauwen L, Hengstler JG, Angrish M, Whelan M. Adverse outcome pathways: a

concise introduction for toxicologists. Arch Toxicol 2017; 1–11 Available from: http://link.springer.com/10.1007/s00204-017-2020-z

21 OECD. AOP definition OECD. Available from: http://www.oecd.org/chemicalsafety/testing/adverse-outcome-pathways-molecular-screening-and-toxicogenomics.htm

22 Krewski D, Acosta D, Andersen M, Anderson H, Bailar JC, Boekelheide K, Brent R, Charnley G, Cheung VG, Green S, Kelsey KT, Kerkvliet NI, Li AA, McCray L, Meyer O, Patterson RD, Pennie W, Scala RA, Solomon GM, Stephens M, Yager J, Zeise L, Stephens M, Yager J, Zeise L. Toxicity testing in the 21st century: a vision and a strategy. J Toxicol Environ Health B Crit Rev 2010; 13: 51–138 Available from: http://www.ncbi.nlm.nih.gov/pubmed/20574894

23 Briggs K, Cases M, Heard DJ, Pastor M, Pognan F, Sanz F, Schwab CH, Steger-Hartmann T, Sutter A, Watson DK, Wichard JD. Inroads to Predict in Vivo Toxicology—An Introduction to the eTOX Project. Int J Mol Sci 2012; 13: 3820–3846 Available from: http://www.mdpi.com/1422-0067/13/3/3820/

24 Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, Evelo CT, Blomberg N, Ecker G,

Goble C, Mons B. Open PHACTS: semantic interoperability for drug discovery. Drug Discov Today 2012; 17 Available from: http://www.sciencedirect.com/science/article/pii/S1359644612001936?via%3Dihub

25 Raies AB, Bajic VB. In silico toxicology: computational methods for the prediction of chemical toxicity. Wiley Interdiscip Rev Comput Mol Sci 2016; 6: 147–172 Available from: http://doi.wiley.com/10.1002/wcms.1240

26 Carrió P, López O, Sanz F, Pastor M. eTOXlab, an open source modeling framework for implementing predictive models in production environments. J Cheminform 2015; 7: 8 Available from: http://www.jcheminf.com/content/7/1/8

27 Cronin MTD, Madden J. In silico toxicology : principles and applications. Royal Society of Chemistry, 2010

28 Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, Light Y, McGlinchey S, Michalovich D, Al-Lazikani B, Overington JP. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012; 40: D1100-7 Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr777

29 Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu

Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y, Wishart DS. DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res 2014; 42: D1091-7 Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3965102&tool=pmcentrez&rendertype=abstract

30 Wang Y, Xiao J, Suzek TO, Zhang J, Wang J, Zhou Z, Han L, Karapetyan K, Dracheva S, Shoemaker BA, Bolton E, Gindulyte A, Bryant SH. PubChem's BioAssay Database. Nucleic Acids Res 2012; 40: D400-12 Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3245056&tool=pmcentrez&rendertype=abstract

31 Dix DJ, Houck KA, Martin MT, Richard AM, Setzer RW, Kavlock RJ. The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals. Toxicol Sci 2007; 95: 5–12 Available from: https://academic.oup.com/toxsci/article-lookup/doi/10.1093/toxsci/kfl103

32 Knudsen TB, Martin MT, Kavlock RJ, Judson RS, Dix DJ, Singh A V. Profiling the activity of environmental chemicals in prenatal developmental toxicity studies using the U.S. EPA's ToxRefDB. Reprod Toxicol 2009; 28: 209–219 Available from:

http://www.ncbi.nlm.nih.gov/pubmed/19446433

33 Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA., Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 2016; 3: 160018 Available from: http://www.nature.com/articles/sdata201618

34 Open PHACTS platform. Available from: https://www.openphacts.org/

35 Berners-Lee T, Hendler J, Lassila O. The Semantic Web. Sci Am 2001; 284: 34–43

36 Antoniou G (Grigoris), Van Harmelen F. A semantic Web primer. MIT Press, 2004

37 RDF. Available from: https://www.w3.org/RDF/

38 RDF Schema. Available from: https://www.w3.org/TR/rdf-schema/

39 OWL. Available from: https://www.w3.org/TR/owl-ref/

40 Masinter L, Berners-Lee T, Fielding RT. Uniform Resource Identifier (URI): Generic Syntax. Available from: https://tools.ietf.org/html/rfc3986

41 Baader F, Calvanese D, McGuinness DL, Nardi D, Patel-Schneider PF. The Description Logic Handbook: Theory, Implementation, and Applications. 2003

42 SKOS. Available from: https://www.w3.org/2004/02/skos/

43 VoID. Available from: https://www.w3.org/TR/void/

44 Allemang D, Hendler J. Semantic Web for the Working Ontologist. 2011

45 Ratnam J, Zdrazil B, Digles D, Cuadrado-Rodriguez E, Neefs J-M, Tipney H, Siebes R, Waagmeester A, Bradley G, Chau CH, Richter L, Brea J, Evelo CT, Jacoby E, Senger S, Loza MI, Ecker GF, Chichester C. The Application of the Open Pharmacological Concepts Triple Store (Open PHACTS) to Support Drug Discovery Research. PLoS One 2014; 9: e115460 Available from: http://dx.plos.org/10.1371/journal.pone.0115460

46 Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, Evelo CT, Blomberg N, Ecker G, Goble C, Mons B. Open PHACTS: Semantic interoperability for drug discovery. Drug Discov Today 2012; 17: 1188–1198 Available from: http://www.ncbi.nlm.nih.gov/pubmed/22683805

47 W3 consortium. SPARQL W3 RECOMMENDATION. Available from: https://www.w3.org/TR/sparql11-overview/

48 Carrascosa MC, Massaguer OL, Mestres J. PharmaTrek: A Semantic Web Explorer for Open Innovation in Multitarget Drug Discovery. Mol Inform 2012; 31: 537–541 Available from: http://doi.wiley.com/10.1002/minf.201200070

49 DTK. GARField. Available from: http://www.openphacts.org/about-open-phacts/what-does-open-phacts-do/garfield

50 BioSolveIT. ChemBioNavigator. Available from: http://www.openphacts.org/about-open-phacts/what-does-open-phacts-do/chembionavigator

51 CNIO. Target dosier. Available from: http://www.openphacts.org/about-open-phacts/what-does-open-phacts-do/target-dossier

52 OPS. Open PHACTS explorer. Available from: http://www.openphacts.org/about-open-phacts/what-does-open-phacts-do/open-phacts-explorer

53 Berthold MR, Cebron N, Dill F, Gabriel TR, Kötter T, Meinl T, Ohl P, Thiel K, Wiswedel B. KNIME - the Konstanz information miner. ACM SIGKDD Explor Newsl 2009; 11: 26 Available from: http://portal.acm.org/citation.cfm?doid=1656274.1656280

54 Warr WA. Scientific workflow systems: Pipeline Pilot and KNIME. J Comput Aided Mol Des 2012; 26: 801–804 Available from: http://link.springer.com/10.1007/s10822-012-9577-7

55 Sanz F, Carrió P, López O, Capoferri L, Kooi DP, Vermeulen NPE, Geerke DP, Montanari F, Ecker GF, Schwab CH, Kleinöder T, Magdziarz T, Pastor M. Integrative Modeling Strategies for Predicting Drug Toxicities at the eTOX Project. Mol Inform 2015; 34: 477–484 Available from: http://doi.wiley.com/10.1002/minf.201400193

56 Obiol-Pardo C, Gomis-Tena J, Sanz F, Saiz J, Pastor M. A Multiscale Simulation System for the Prediction of Drug-Induced Cardiotoxicity. J Chem Inf Model 2011; 51: 483–492 Available from:

http://www.ncbi.nlm.nih.gov/pubmed/21250697

57 Ravagli C, Pognan F, Marc P. OntoBrowser: a collaborative tool for curation of ontologies by subject matter experts. Bioinformatics 2017; 33: 148–149 Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw579

58 Codd E., Codd S., Salley C. Providing OLAP to User-Analysts: An IT Mandate. 1993 Available from: http://www.minet.uni-jena.de/dbis/lehre/ss2005/sem_dwh/lit/Cod93.pdf

59 López-Massaguer O, Sanz F, Pastor M. An automated tool for obtaining QSAR-ready series of compounds using semantic web technologies. Bioinformatics 2017; Available from: http://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btx566/4107533/An-automated-tool-for-obtaining-QSARready-series

60 Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Adv Drug Deliv Rev 2012; 64: 4–17 Available from: http://www.ncbi.nlm.nih.gov/pubmed/11259830

61 PostgreSQL database system. Available from: http://www.postgresql.org/

62 OpenPHACTS API. Available from: https://dev.openphacts.org/

63 Odersky M, Spoon L, Venners B. Programming in Scala. Artima, 2008

64 Dalby A, Nourse JG, Hounshell WD, Gushurst AKI, Grier DL, Leland BA, Laufer J. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited. J Chem Inf Model 1992; 32: 244–255 Available from: http://pubs.acs.org/cgi-bin/doilookup/?10.1021/ci00007a012

65 Hilton P, Bakker E, Canedo F. Play for Scala: Covers Play 2. Greenwich, CT, USA: Manning Publications Co., 2013

66 Oracle. JDK8. 2014; Available from: https://www.java.com/

67 RDKit Chemoinformatics toolkit. Available from: http://www.rdkit.org

68 Parsons S. A Semantic Web Primer. 2009

69 Agrawal R, Gupta A, Sarawagi S. Modeling

multidimensional databases. In: Proceedings 13th International Conference on Data Engineering. IEEE Comput. Soc. Press, 1997: 232–243 Available from: https://dl.acm.org/citation.cfm?id=653299

70 Yap CW. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J Comput Chem 2011; 32: 1466–1474 Available from: http://www.ncbi.nlm.nih.gov/pubmed/21425294

71 Carbonell P, Lopez O, Amberg A, Pastor M, Sanz F. Hepatotoxicity prediction by systems biology modeling of disturbed metabolic pathways using gene expression data. ALTEX 2017; 34: 219–234 Available from: http://www.ncbi.nlm.nih.gov/pubmed/27690270

72 Jupp S, Burdett T, Malone J, Leroy C, Pearce M, Mcmurry J, Parkinson H. A New Ontology Lookup Service at EMBL-EBI. Available from: http://ceur-ws.org/Vol-1546/paper_29.pdf