Next Generation of Informatics Tools for Big Data Analytics in Drug Discovery

María Carmen Carrascosa Baena

TESI DOCTORAL UPF / 2017

DIRECTOR DE LA TESI

Dr. Jordi Mestres

DEPARTAMENT CEXS

upf. Universitat Pompeu Fabra Barcelona

A mi familia/ohana

I've seen things you people wouldn't believe.

Attack ships on fire off the shoulder of Orion.

I watched C-beams glitter in the dark
near the Tannhäuser Gate.

All those moments will be lost in time,
like tears in rain.

Time to die.

Roy Batty
Nexus 6 replicant
serial code N6MAA10816
Blade Runner

This is true love. You think this happen every day?

William Goldman, The Princess Bride

Summary

The classical silver bullet paradigm of one drug interacting with a single target linked to a disease is currently challenged. It is now widely recognized that one drug interacts with multiple targets and these targets are involved in many biological pathways and expressed in a variety of organs. As the notion of complexity has been gradually accepted, the reductionist drug discovery approach has naturally evolved towards systems multilevel strategies.

Thanks to technological advances, there has been a huge increase of data generated in the various fields relevant to drug discovery, namely, chemistry, pharmacology, toxicology, genomics, metabolomics, etc., which has expanded dramatically our ability to generate computational models with increasing performance and coverage. But ultimately, extracting knowledge from this complex, vast and heterogeneous amount of data is not straightforward.

The main objective of this Thesis is to develop new interactive analytics and visualization tools and investigate their ability to extract knowledge from highly interconnected data when implemented into an integrated flexible platform to facilitate drawing simple answers from complex questions. In particular, special emphasis will be put in the navigation aspects of the relationships between systemic entities (small molecules and their metabolite, protein targets, safety terms).

Resum

El paradigma clàssic on un medicament interacciona amb un únic target biològic vinculat a una malaltia es posa en dubte. Actualment es reconeix que un medicament interacciona amb múltiples targets biològics i que aquests targets estan involucrats en multitud de pathways i que s'expressen en una varietat d'òrgans. Amb el creixent reconeixement d'aquesta complexitat, la estratègia reduccionista del procés de descoberta de nous medicaments ha evolucionat cap a estratègies sistèmiques multinivell.

Gràcies als avenços tecnològics, hi ha hagut un gran increment de les dades generades en les diverses àrees rellevants en la descoberta de nous medicaments: química, farmacologia, toxicologia, genòmica, metabolòmica, etc fet que ha expandit considerablement la nostra habilitat per general models computacionals amb un rendiment i cobertura creixents. Però darrerament, extreure coneixement d'aquest complex, vast i heterogeni volum de dades no és simple.

El principal objectiu d'aquesta tesi es desenvolupar noves eines analítiques i de visualització i investigar la seva capacitat per extreure nou coneixement de dades altament interconnectades; eines integrades a una plataforma flexible que per obtenir respostes simples a preguntes complexes. En particular, farem èmfasi en la navegació per les relacions entre les entitats del sistema (molècules petites i els seus metabòlits, proteïnes com a targets biològics, termes de safety).

Preface

This Thesis started at the Research Group on Systems Pharmacology, within the Research Program of Biomedical Informatics (GRIB) of the University Pompeu Fabra and the IMIM Hospital del Mar Medical Research Institute, and it was completed at Chemotargets, the spin-off company the group.

From the very beginning, the main goal was to explore and develop graphical tools to navigate in highly interconnected and heterogeneous data, that can be used by academic, corporate, and non-for-profit organizations but taking into consideration that the main potential end user may likely be a high-level scientist or decision-makers that usually are not computer experts. Therefore, focus was given to develop easy to use, but at the same time, powerful graphical environments focused on the specific needs and questions to be addressed. With this in mind, the right approach is to integrate data visualization in the context of a highly interactive user interface tools.

Visualization tools have been used for a long time to get insights from data in a more user-friendly manner by optimally integrating information retrieval and data visualization per se. This Thesis presents our evolving efforts towards developing what is currently considered one of the most powerful analytics and visualization platforms for pharmacology and safety profiling of small molecules.

Table of contents

1 INTRODUCTION

1.1 Big Data, Big Challenge

When you like a page on Facebook, or buy the last best seller on Amazon or create a playlist in Spotify or when you drive with the location mode on in your mobile, you are generating digital data that is going to make part of this large collection of information commonly known as Big Data.

Digital data is doubling in size every 2 years and will multiply 10-fold between 2013 and 2020 – from 4.4 trillion gigabytes to 44 trillion gigabytes [1].

The term Big Data has been used since 1990's but was Doug Laney, an analyst at Gartner, who in February 2001 published a research note [2] where he defined the main characteristics of Big Data, the 3 V's: volume, velocity, and variety. The volume describes the amount of data, velocity describes the frequency at which this data is generated and variety refers to the data structure or format, from excel spreadsheets, photo, audio, video, GPS data, sensor data, documents, etc. Orbiting this 3 original, new V's have been added over the last few years: variability, veracity, visualization… This article [3] summarizes all the V's added to define Big Data and also make his own approach with 42 V's.

As we can see, there are many descriptions of the characteristics of big data but despite the wide use of the term, there is not a consensus about its definition. One of the most quoted definitions was the one provided by the McKinsey Global Institute in 2011: “Big data” refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze. This definition is intentionally subjective and incorporates a moving definition of how big a dataset needs to be in order to be considered big data—i.e., we don’t define big data in terms of being larger than a certain number of terabytes (thousands of gigabytes). We assume that, as technology advances over time, the size of datasets that qualify as big data will also increase. Also note that the definition can vary by sector, depending on what kinds of software tools are commonly available and what sizes of datasets are common in a particular industry. With those caveats, big data in many sectors today will range from a few dozen terabytes to multiple petabytes (thousands of terabytes) [4].

Summarizing, we are talking about a big amount of complex and heterogeneous data that grow very quickly. The question is what we can do with these data? How can we use it? Can we take some benefit from it?

We can find several real examples of Big Data methods in real business: Google search engine [5], Google AdSense [6], Google AdWords platform [7] and in general most of the web businesses based on analysis of users behaviour.

For example, the Google advertising platform AdWords, analyses user behavior data for tailoring the advertising we see.

Amazon or Spotify use their data to create recommendations to the users based on previous purchases or our listening habits.

All these platforms rely on a high computational power to process a high amount of data combined with mathematical models (PageRank, Deep Learning, etc) to obtain better predictions like the most suitable advertisement or cross-selling recommendations.

We can also find Big Data examples in the scientific research. For example, the ATLAS (Argonne Tandem Linear Accelerator System) [8] experiment performed at the Large Hadron Collider (LHC) at CERN in Geneve is a paradigmatic example of the application of Big Data principles and methods to scientific research.

The LHC experiments generate 100 PB per year of raw data that have to be stored and analyzed in an efficient way. The storage of this amount of data was a big challenge solved with ROOT [9] a framework for data processing born at CERN, that is extremely powerful for fast access to huge amounts of data. Physicists around the world access all these data. This means a considerable movement of data (several petabytes per week) that requires tools like Big PanDA [10] that provides a very large scale data-intensive distributed computing.

One of the most important results of the analysis of these data is the confirmation of the Higgs boson [11]. In July 2017, CERN confirmed that all measurements still agree with the predictions of the Standard Model, and called the discovered particle simply "the Higgs boson" [12].

Another example, in the area of genomics, is the 1000 Genomes Project [13], started in 2008 with the objective to develop an extensive catalog of variation in the human genome by sequencing the genomes of at least 1000 individuals from around the world. In March 2012, the still growing project resources include more than 260 terabytes of data in more than 250.000 publicly accessible files. New methods for data submissions and access were developed in the project [14] to make all of these data available to the researcher's community. As a result, many studies have been published using this data [15]. Other examples include Genome-wide Association Studies (GWAS), filtering non-pathogenic variants from exome, whole genome and cancer genome sequencing projects, and genetic analysis of population structure and molecular evolution, among others.

1.2 Data-driven or Hypothesis-Driven

But the use of big data in science has some controversy.

The controversy started with the publication of “The end of theory”[16] where the author defends the idea that the classic scientific hypothesis-driven method was obsolete because the big amount of available data represents a new way to make science. This is what we understand as data-driven science. Some responses appear against this argument defending the classical hypothesis-driven method.

We can consider the use of big data in the LHC exposed in the previous section as a case of hypothesis-driven method. The importance of the demonstration of the Higgs boson existence, already theorized in the 1960, led to the construction of the CERN’s Large Hadron Collider in attempt to create Highs boson and other particles for observation and studies. The use of all the information generated by the LHC was used to confirm the hypothesis of the existence of the Higgs boson.

On the other hand, we can consider the 1000 Genome Project applications as data-driven studies. First, all the data was collected and made available to the scientific community. For sure, some of the studies of these data started with some kind of background hypothesis but the main idea was to explore the data to extract some insight and build new hypothesis from the data.

However, despite these discussions, even the rejecters of the data-driven argument agree that we have to take profit of this amount of data, keeping always in mind that big data does not necessarily mean big information. In fact, distilling an informative data puddle from a big data sea is nowadays a highly active area of research as it has important implications for the design and performance of big data analytics and visualization tools.

In particular, in the area of drug discovery, apart from this methodological dispute, the use of big data has to deal with two main issues: the heterogeneity of the data (from disciplines as diverse as chemistry, both medicinal and analytical, pharmacology, toxicology, and genomics, to name just a few) and the dispersion of the data in different data silos that may be geographically remote. To solve these problems many strategies have been proposed. One of them was Open PHACTS [17], a 5-year project funded by a European grant from the Innovative Medicines Initiative (IMI), ending in February 2016. The project was born as a public–private partnership between academia, publishers, small and medium-sized enterprises and pharmaceutical companies. The main goal of the project was the creation of a freely available platform, integrating data from a variety of information resources, and providing tools and services to query these integrated data to support drug discovery research.

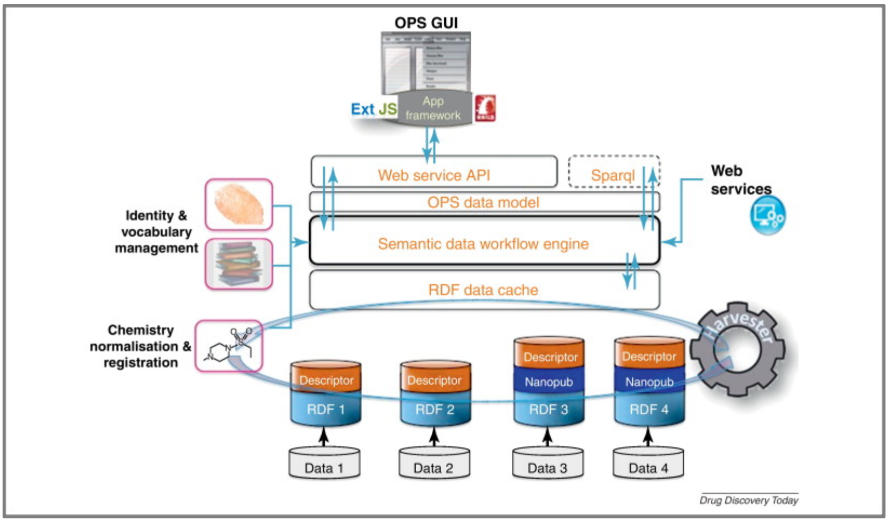


Figure 1:The OPS platform architecture[17].

As part of the project, some example applications were developed to illustrate how the Open PHACTS platform can be used to answer diverse research questions in the field of drug discovery. One of these applications was Pharmatrek [18], an integrative and interactive web application that allows scientists to extract new knowledge from the Open PHACTS platform and that is part of this Theses (section 3.2). The main goal was to provide visual tools that allow the user to define custom questions on the interaction between drugs and targets. Pharmatrek was the conceptual and technical foundation that would later lead to the development of CT-link (section 3.3) and, more recently, to CLARITY (section 3.4), one of the most powerful data analytics and visualization tools available today in preclinical drug discovery.

1.3 The need and purpose of data visualization

Projects like Open PHACTS solve partly the problem of data access by integrating different data sources in a single repository (post-project management and maintenance issues aside), but they do not provide advanced analytics tools for the exploration of the data, a key aspect in data-driven drug discovery.

To really get a deeper insight into these data, visualization tools are essential because of the amount and complexity of the data involved. But this is not as new as we may think. Data visualization has been used for centuries to represent complex data in a more accessible way that helps the user to understand and extract information from the data.

For example, the famous Majorcan philosopher Ramon Llull (1232-1315) used data visualization to represent a unified vision of the knowledge as trees. The roots are the principles, the trunk is the structure and the branches the genres. This illustration is included in his book entitled *Arbor scientiae*, published around 1295[1]. This knowledge tree reminds the use

[1] http://bibliotecadigital.rah.es/dgbrah/es/consulta/registro.cmd?id=44591

of ontologies [19] in modern knowledge representation and semantic web technology.

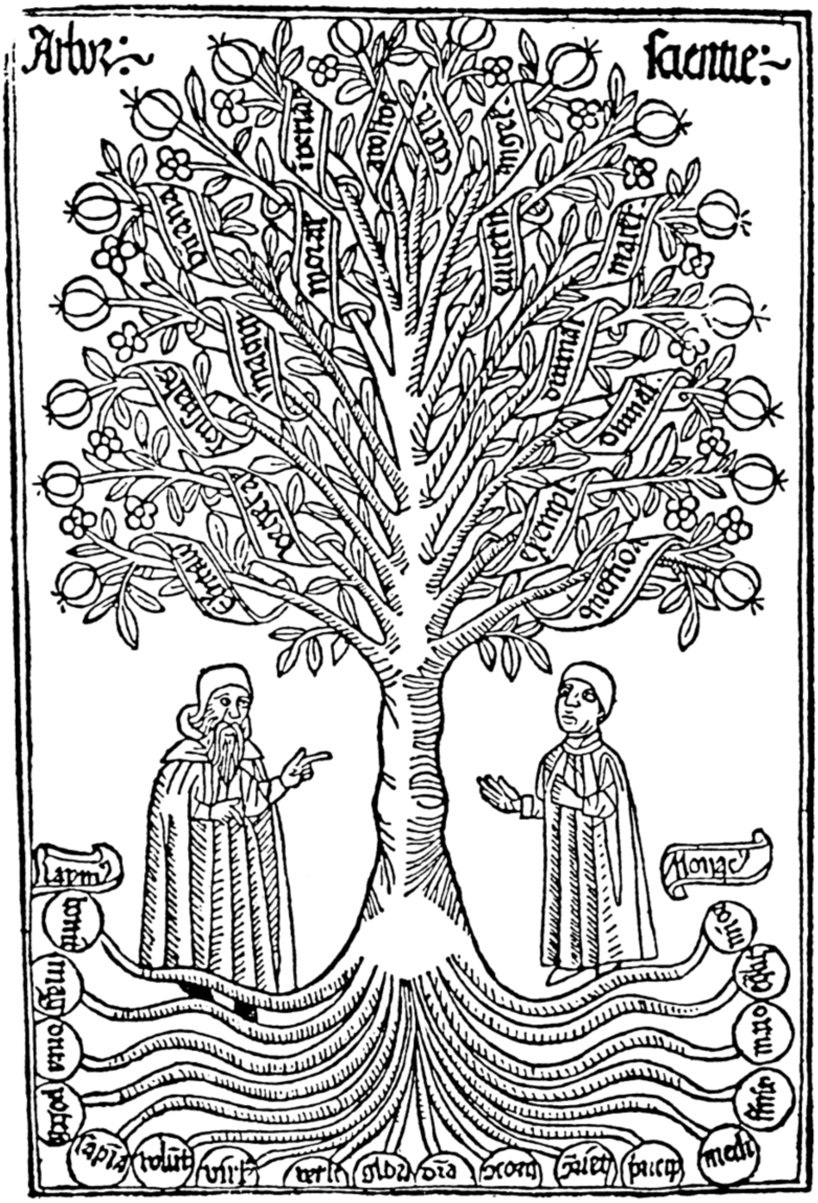


Figure 2: Ramon Llull's representation of knowledge.

In fact, most of our current widely used statistical charts, such as the line chart, bar chart or pie chart, are not recent creations. They were created in the 18th century by William Playfair (1759-1823) a Scottish economist. His book *The Commercial and Political Atlas* published in London in 1786 is considered the first major work to contain statistical graphs.

We can also find a reference on the use of data visualization to support scientific discovery in the mid 19th century. At this time, London was gripped by cholera. According to the doctors, the disease was spread by "miasma" in the air. But Dr. John Snow hypothesized that the source of cholera spread could be through contaminated water. In 1854, the cholera was focused in the Soho district and to proof his theory Dr. Snow mapped the 578 known cholera deaths around Soho in a map of the area and then add the location of the 13 public water pumps in the same area. The map revealed an amazing pattern, leading to a spatial clustering of cases around a water pump on Broad Street. Dr. Snow took then some water samples from the area and confirmed the presence of an unknown bacterium in the samples. He had the pump handle removed, and the cholera outbreak quickly subsided. This case is a paradigmatic example of how visualization helps to identify patterns ("cases around a water pump") and formulate a hypothesis ("the cause is related to the water source"). This event marked one of the first cases of modern epidemiological research.

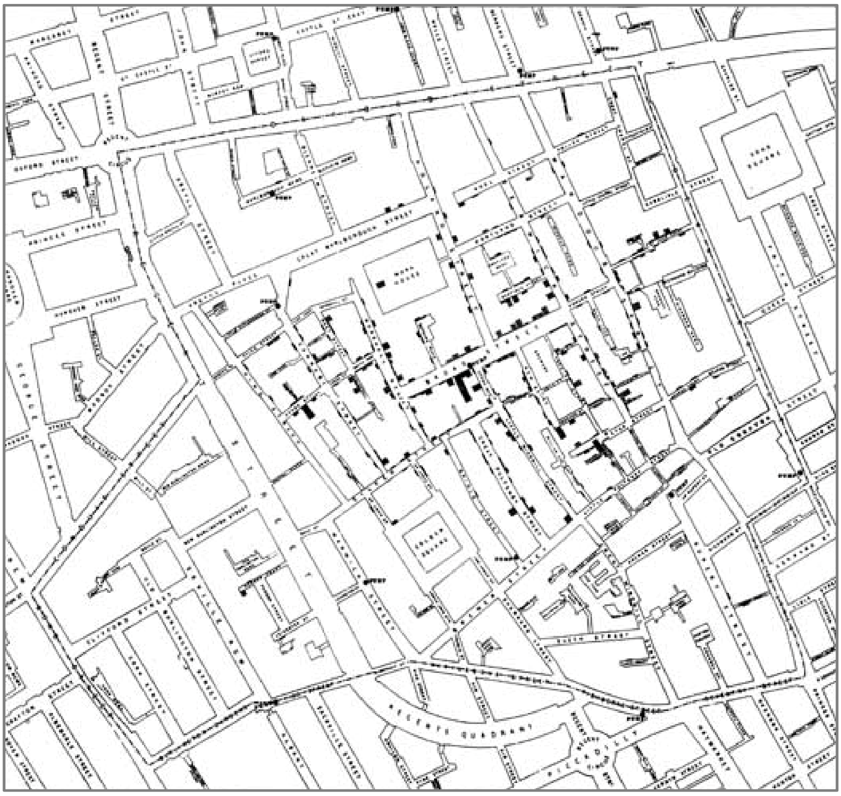


Figure 3: Dr. John Snow cholera map.

Moving to the 20th century, we must highlight the figure of John Tukey, considered the father of exploratory data analysis. John Tukey was an American mathematician that worked at Princeton University and Bell Telephone Laboratories. He made great contributions to the foundations of statistics, the analysis of variance, time series, and inferences involving multiple parameters.

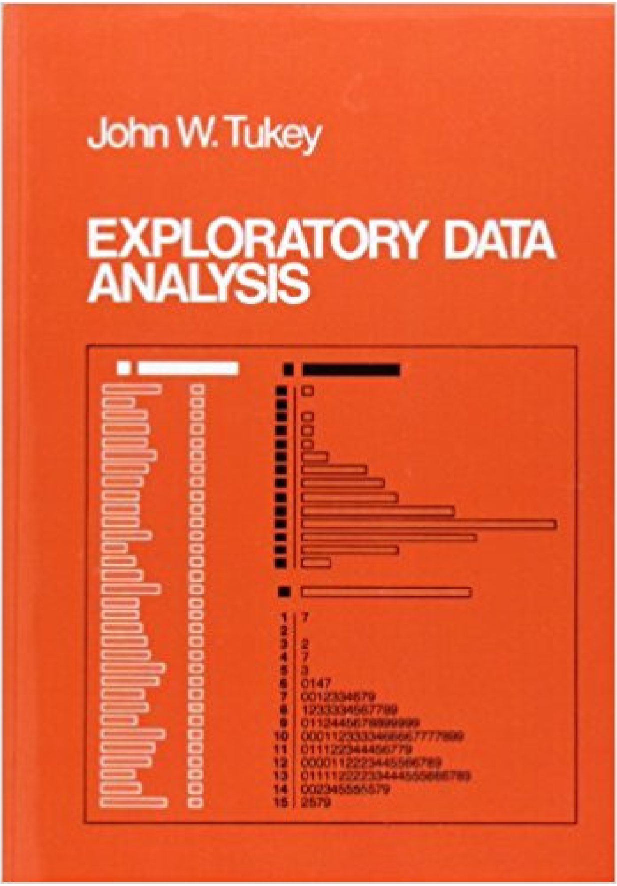


Figure 4: Tukey "Exploratory data analysis" cover.

In his book “Exploratory data analysis” [20], Tukey introduces the idea of using statistics not just to confirm hypotheses but to suggest new ones. The idea is to collect data and apply exploratory methods to form preliminary hypothesis. In this analysis of the data, Tukey makes a special emphasis in the use of visual methods. In one article published in 1974 [21], Tukey expresses clearly why data visualization is key to data exploration:

“Why do we use pictures? Most crucially to see behaviour we had not explicitly anticipated as possible—for what pictures are best at is revealing the unanticipated; crucially, often as a way of making it easier to perceive and understand things that would otherwise be painfully complex. These are the important uses of pictures.”

Which in summary means that “the main tasks of pictures are to reveal the unexpected and to make the complex easier to perceive”.

Figure 5: Tukey showing PRIM-9 system.

In a historical video that is available on YouTube (https://www.youtube.com/watch?v=B7XoW2qiFUA), one can see Tukey introducing the properties and uses of PRIM-9[22], the first program to use interactive, dynamic graphics for viewing and dissecting multivariate data. This was in 1974 and it is quite amazing to see how the software presented is not that far away from some of the current statistics software.

Today a major figure in the field of data visualization is Edward Tufte, Professor Emeritus of Political Science, Statistics, and Computer Science at Yale University. His research concerns statistical evidence and scientific visualization. Tufte has published 4 books until now, the contents of which is still being regarded as the main reference for data visualization. Of mention is the fact that the first of those books, entitled “The Visual Display of Quantitative Information”, was published in 1983.

For Tufte, graphics are not just useful for displaying numbers but for clarifying just about anything one person is trying to tell someone else, putting emphasis on the communication aspect of data visualization. It is all about the relationship between the viewer and the information on the screen, and the viewer's cognitive tasks in looking at that information.

Tufte agrees that the digital world has opened up more possibilities with visualization, making a particular remark on the aspect that the principles for showing information do not depend (should not be dependent) on the platform or the media used.

For example, Tufte considers that Feynman diagrams are one of the best visualization tools having ever been created. They first appeared in a 1949 paper [23], much before any computer graphics had emerged.

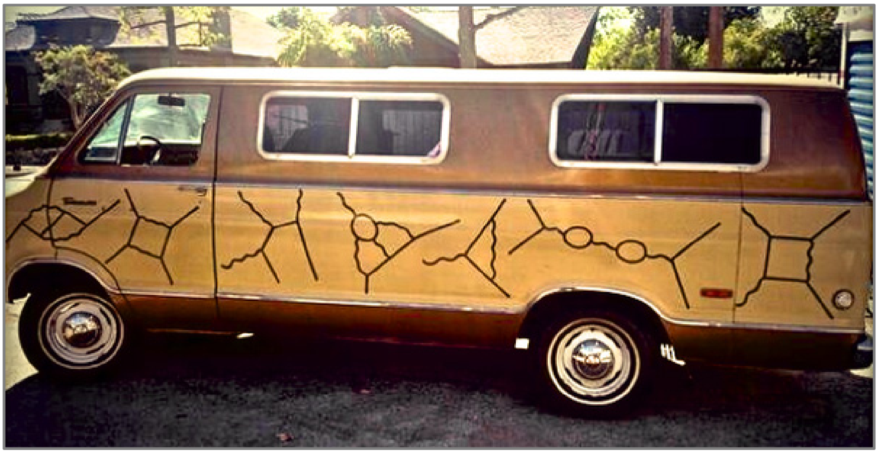


Figure 6: Richard Feynman van with Feynman diagrams.

Feynman diagrams are simplified visual representations of particle interactions in quantum mechanics that have been used for 70 years by scientists. Frank Wilczek, the Nobel Prize laureate in Physics of 2004, said once that the calculations that eventually got him to the Nobel Prize would have been literally unthinkable without the use of Feynman diagrams, as would all calculations that established a route to production and observation of the Higgs particle. This is a clear example that a rather simple visualization may be of enormous help in a very complex problem.

1.3.1 DATA VISUALIZATION PRINCIPLES

For Tufte, a good visualization consists of complex ideas communicated with clarity, precision, and efficiency.

In his books, Tufte suggests 6 principles to design good data visualization tools:

- Focus on content: The focus should be on the content of the data, not the visualization technique. This leads to using a clear, simple, straightforward design with a richness of data.

- Show comparisons: To put data into context or simply to know where your current data sits relative to some data references

- Show causality

- Use multivariate data

- Complete integrated nodes: To improve displays of data, it is important to completely integrate text, images and numbers. It is also important to include any information or data that is relevant. This is particularly important in the inclusion of explanatory text on figures

- Establish credibility: Tufte argues repeatedly that including source information is one of the most important aspects of creating a convincing visual display or a convincing presentation. Allowing the viewers to access the source material will give them confidence in your result

1.3.2 INTERACTIVE COMPUTING SYSTEMS

To wrap up this introduction, I believe it is important to stress one final point. Without applying advanced human-computer interaction techniques, such as the modern graphical user interface (GUI), the ideas exposed in the previous section could not be implemented in practical tools.

During the first years of computation, computers were enormous machines (basically requiring a full room) with a big calculation power. Apart from the input process, we cannot say that these computers were very interactive. The user entered some input and after some hours (often days and weeks) of calculations an output was provided.

In this respect, researchers like Douglas Carl Engelbart were pioneers in the field of human-computer interaction. Engelbart was an American engineer and inventor who worked on establishing the fields of human-computer interaction in the Augmentation Research Center at Stanford Research Institute in Menlo Park, CA.

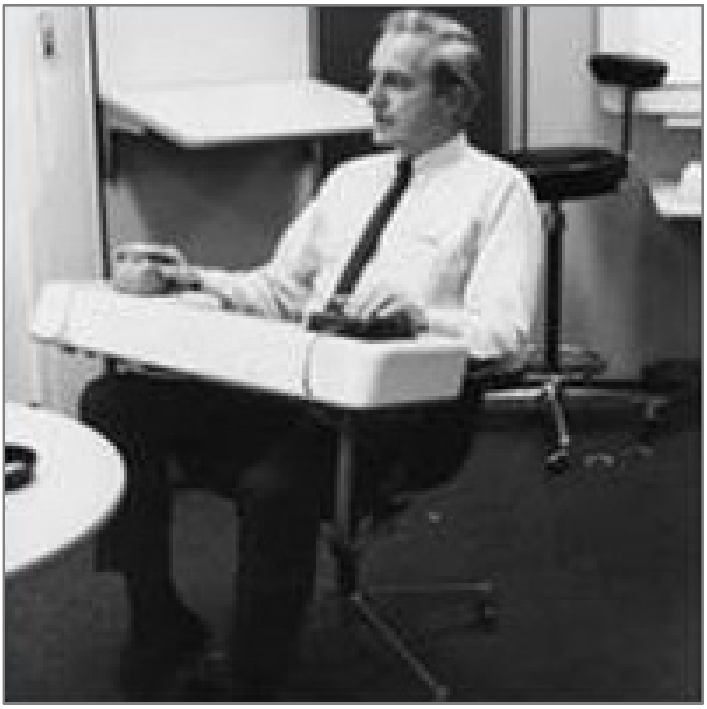


Figure 7: Douglas Engelbart using a mouse.

In December 1968, he made a presentation, known as the "the mother of all demos", at the Fall Joint Computer Conference in San Francisco. For the event, in front of over a thousand computer scientists, he sat on stage in front of a mouse, a keyboard, and other controls and projected the computer display onto a 22-foot-high video screen behind him.

In a live demo of 90 minutes, which I highly recommend to watch (https://www.youtube.com/watch?v=VScVgXM7lQQ), he presented the oN-Line Systems (NLS) that included pretty much all fundamental elements of modern computing systems: mouse (which he invented in 1965), windows, hypertext, graphics, navigation, command input, networking, video conferencing, word processing, file linking and revision control, all in a single system.

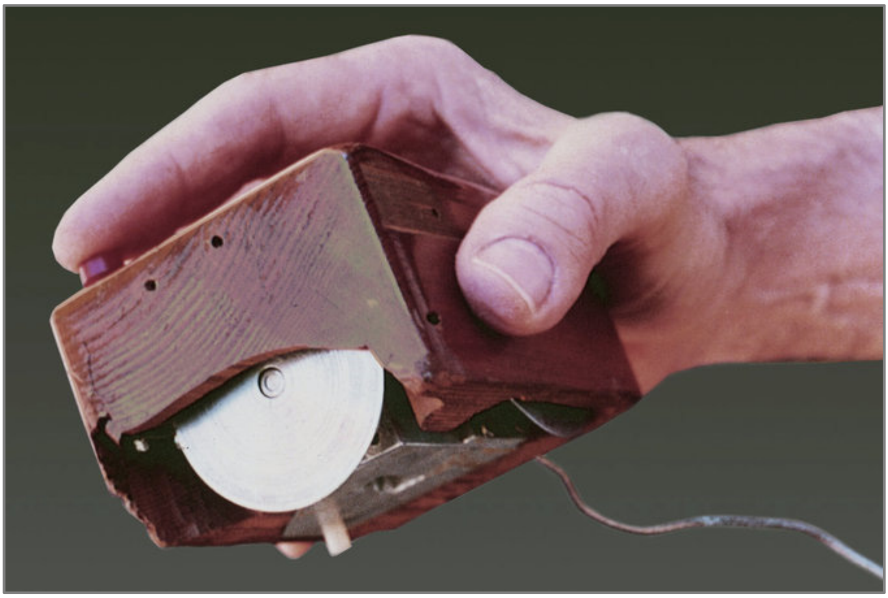


Figure 8: First computer mouse from 1964.

These were the very first steps of a GUI development, a technology that is key for human-computer interaction and that nowadays is part of every piece of modern computing platforms, from smartphones to tablets, laptops, desktop computers and even supercomputers, something that helps the user to interact more easily with a computerized system.

1.4 Brief review of the current state of the art in data visualization tools

In the previous sections, we have seen the importance of interaction and data visualization for the exploration of data. Now we are going to make a brief review of the available software or applications in the area of drug discovery that take into account these two basic concepts: visualization and interaction.

For example, there are some general softwares, like Tableau [24] or TIBCO Spotfire [25] (proprietary), that provide general tools to integrate queries and navigate data.

Tableau started like an academic project at Stanford University and became a powerful business intelligence software platform. Despite that fact that Tableau is a very powerful tool for data visualization, it is a general-purpose software application not particularly focused on drug discovery. The fact that the software is general offers the possibility to use it in different fields, which can be a positive quality at first. But it can be also a disadvantage when you want to apply it in a more specific field such as drug discovery. To use the software, you have to develop an infrastructure around it, to provide the data needed, and to specify the outputs that you want to show. You need to be quite an expert to adapt the application to your needs, as some custom developments have to be made.

In the same category, TIBCO Spotfire is a very powerful general-purpose software that has to be adapted to solve your specific problem. Usually, this adaptation is not trivial and implies a substantial cost in time and money.

On the other hand, some more specific softwares for drug discovery also exist. DataWarrior [26] is a good example.

DataWarrior is an interactive data analysis and visualization software with some sort of chemical intelligence, which was developed in 2014 by Actelion Pharmaceuticals Ltd. The tool is focused on the exploration of the chemical space but it does not provide out-of-the-box tools to explore chemical data from a biological perspective. The fact that it is a desktop application can be considered for many a drawback as it does not work in web environments.

Another major group of tools are those developed around public data repositories, like ChEMBL [27], PubChem [28], DrugBank [29], or ChemSpider [30].

Although they provide simple and useful web applications, they do not offer the possibility to ask complex questions and explore and query big heterogeneous data. Also, they do not provide mechanisms of data integration of different sources beyond the record to record cross-referencing.

Therefore, if complex queries and analyses are needed between different sources of heterogeneous data, the researcher is forced to download all data from the different repositories and build custom tools to perform data integration, querying, and analysis. The recently developed Open PHACTS platform partly addresses some of these limitations.

With all of this in mind, the main idea of this thesis is to take profit of the highly interactive power provided by current computers and devices. We will combine this interactivity with modern data visualization techniques to create specific tools to explore pharmacological data in a unique highly interactive way that allow the user to interrogate the data with enough freedom/flexibility to address questions that lead to smart answers and create new knowledge. All tools were developed on the solid foundation of the predictive models developed by Chemotargets over the last decade.

2 OBJECTIVES

The main objective of this PhD Thesis is to develop novel analytics and visualization tools that allow for exploring pharmacology and safety data in a highly interactive and friendly manner leading to the generation of new knowledge.

The specific objectives are:

- Develop a first generation tool based on a web application (Pharmatrek) that allows the user explore data from the Open PHACTS project in an interactive way and connecting with other web applications developed in the project.

- Create a second generation integrated platform for drug discovery (CT-link) that integrates the PredictFX software developed by Chemotargets and a graphical user interface (GUI) and allows the user to execute the prediction components and navigate the results in an interactive and flexible way.

- Develop a third generation of highly interactive web application (CLARITY) that allows scientists to ask complex drug discovery questions in a highly interactive and visual environment. The tool will allow the user to explore the complex connection between the various entities of therapeutic relevance (small molecules, metabolites, proteins and safety terms) and establish and understand the potential relationships between them (molecule-metabolite, molecule-protein, molecule-safety, metabolite-protein, metabolite-safety, protein-safety). The new platform will provide tools to search, filter, sort and group across all entities. It will also provide analytics and visualization tools to facilitate the connection and navigation across highly heterogeneous big data.

3 RESULTS

3.1 Introduction

This thesis started at the System Pharmacology Group that is part of the Research Programme of Biomedical Informatics (GRIB). The group aims to develop and apply computational tools for the systematic identification of active molecules for therapeutically relevant target families to be used either upstream as chemical probes for target validation or downstream as hits for lead generation within the drug discovery process.

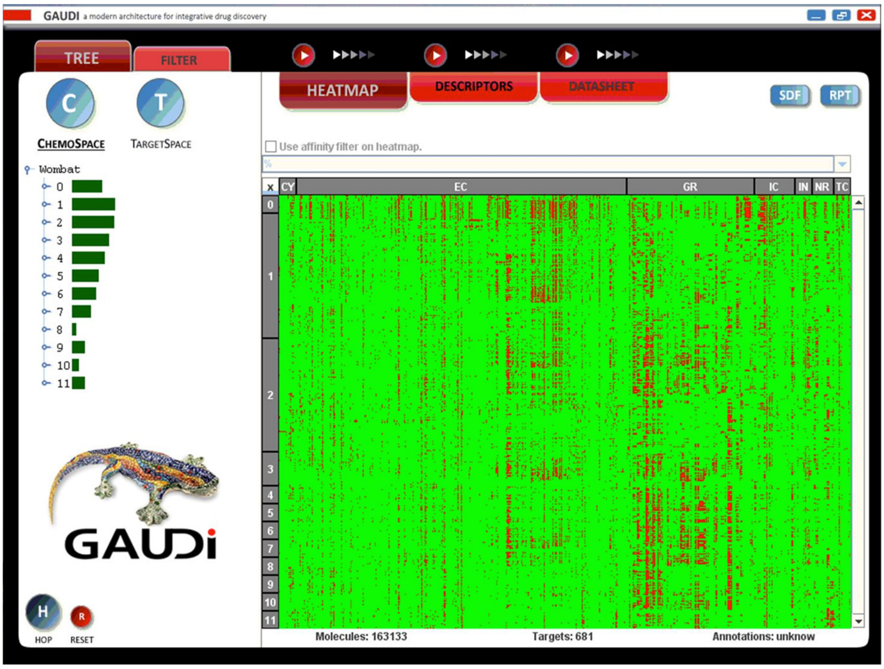


Figure 9: Gaudi platform main page.

To achieve this goal, the group has been exploring and gaining experience on the potential power of the use of visualisation tools over a decade. Some early examples of these efforts are the development of the iPhace platform [31] and the first technology transfer attempt with the GAUDI platform (Figure 9).

When I joined the research group, my first task was to develop web services for the EU-ADR project [32]. This was a data integration project to provide scientists with a one-stop-shop platform to access data on adverse drug events.

The next project I was assigned to was the development of the BlastXP [33] web application. This project started as a synergistic collaboration inside the group. Ferran Briansó, another member of the group, developed a methodological framework for assessing the cross-pharmacology between targets. At the same time, I was asked to re-design the web applications of the group. Therefore, the idea to create a web application to provide easy access to cross-pharmacology data emerged naturally. Both the EU-ADR and the BlastXP projects gave a solid basis to take on the development of a novel integrative platform to navigate on both pharmacology and safety data.

That opportunity came when the research group participated in the Open PHACTS project [17] and we started the development of the Pharmatrek application [18]. It just happened that Chemotargets, the spin-off company created in

2006 from the research group, was also participating in the Open PHACTS project.

When the Open PHACTS project ended, I joined Chemotargets to design a new integrative platform for drug discovery based on the experience gained in the development of Pharmatrek. This would be the origin of CT-link, Chemotargets first software product with an easy-to-used graphical user interface (GUI) in the field of predictive pharmacology and safety. The platform was well appreciated by pharmaceutical companies and academic institutions and Chemotargets distributed 9 licenses worldwide, covering markets from Europe to Japan.

Finally, the experience gained in the development of CT-link led to the design of a new generation of data analytics and visualization platform called CLARITY. The official date for the launch of version 1 of this new software product from Chemotargets is October 2nd, 2017. Following the strategic investment agreement with the Prous Institute of Biomedical Research on May 2017, current expectations for the adoption of CLARITY by pharmaceutical companies are very high and before launch we have already 11 additional requests for testing, over and above the transition of the previous 9 licenses from CT-link to the new platform. This is a wonderful example of a successful technology transfer initiative from an academic idea to a commercial product and I am delighted to have been a part of it.

3.2 Review of the results

This thesis presents the path towards developing what is currently considered the state-of-the-art platform in data analytics and visualization of pharmacology and safety data:

- Development of Pharmatrek, a first generation of visualization platform: it is a web application developed under the remit of the Open PHACTS[17] platform providing unified data access through Open PHACTS API [34]. Pharmatrek constituted my first attempt to visualize pharmacological data in a highly interactive manner. The following aspects had to be considered during the duration of the project:
    - To modify the interface to meet evolving user needs and requirements and to adapt data access to the new versions of the Open PHACTS API and connection with other applications in the project.
    - During the last year of the project, we started a collaboration with the company Chemotargets to adapt the tool to visualise the data generated by the PredictFX software [35], at the time being distributed by CERTARA.
    - See 3.5 for more details.

- Development of CT-link, a second generation of visualization platform: Develop a visualisation platform to analyze data generated by the PredictFX software from Chemotargets:
  - Convert PredictFX from a command line software to a web application.
  - Different pharma companies acquired the resulting CT-link software.
  - See 3.6 for more details
- Development of CLARITY, a third generation of visualization platform:
  - Due to the announced end of life of Adobe Flex, on which CT-link was based, a new design was necessary to ensure the sustainability of the CT-link platform.
  - The incorporation of new entities: metabolites and neighbours.
  - Development of new analytics and visualization tools to explore the complex network of interactions between all entities involved: small molecules (including metabolites and neighbours), protein targets and safety events.

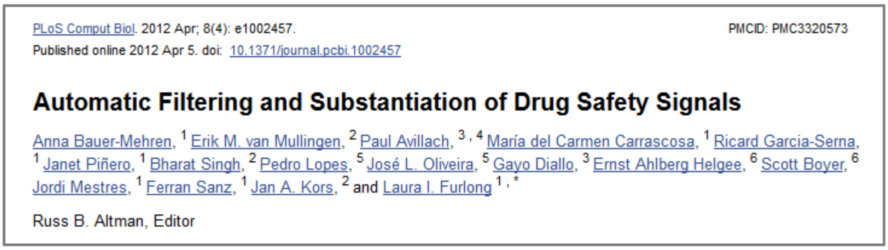
- To prepare the software architecture to support the future increase on the number of entities (pathways, organs, diseases, …) and relations between them.

- To implement new tools to explore the existing known space, the data background used for building the predictive models.

- Use html5 to make it responsive and avoid the use of Adobe Flash as in CT-link.

- See 3.7 for more details.

3.3 Papers EU-ADR Project



PLoS Comput Biol. 2012 Apr; 8(4): e1002457. PMCID: PMC3320573
Published online 2012 Apr 5. doi: 10.1371/journal.pcbi.1002457

Automatic Filtering and Substantiation of Drug Safety Signals

Anna Bauer-Mehren,[1] Erik M. van Mullingen,[2] Paul Avillach,[3,4] María del Carmen Carrascosa,[1] Ricard Garcia-Serna,[1] Janet Piñero,[1] Bharat Singh,[2] Pedro Lopes,[5] José L. Oliveira,[5] Gayo Diallo,[3] Ernst Ahlberg Helgee,[6] Scott Boyer,[6] Jordi Mestres,[1] Ferran Sanz,[1] Jan A. Kors,[2] and Laura I. Furlong[1,*]

Russ B. Altman, Editor

Abstract: Drug safety issues pose serious health threats to the population and constitute a major cause of mortality worldwide. Due to the prominent implications to both public health and the pharmaceutical industry, it is of great importance to unravel the molecular mechanisms by which an adverse drug reaction can be potentially elicited. These mechanisms can be investigated by placing the pharmaco-epidemiologically detected adverse drug reaction in an information-rich context and by exploiting all currently available biomedical knowledge to substantiate it. We present a computational framework for the biological annotation of potential adverse drug reactions. First, the proposed framework investigates previous evidences on the drug-event association in the context of biomedical literature (signal filtering). Then, it seeks to provide a biological explanation (signal substantiation) by exploring mechanistic connections that might explain why a drug produces a specific adverse reaction. The mechanistic connections include the activity of the drug, related compounds and drug metabolites on protein targets, the association of protein targets to clinical events, and the annotation of proteins (both protein targets and proteins associated with clinical events) to biological pathways. Hence, the workflows for signal filtering and substantiation integrate modules for literature and database mining, in silico drug-target profiling, and analyses based on gene-disease networks and biological pathways. Application examples of these workflows carried out on selected cases of drug safety signals are discussed. The methodology and workflows presented offer a novel approach to explore the molecular mechanisms underlying adverse drug reactions.

The EU-ADR Web Platform: delivering advanced pharmacovigilance tools

José Luis Oliveira ✉, Pedro Lopes, Tiago Nunes, David Campos, Scott Boyer, Ernst Ahlberg, Erik M. van Mulligen, Jan A. Kors, Bharat Singh, Laura I. Furlong, Ferran Sanz, Anna Bauer-Mehren, Maria C. Carrascosa, Jordi Mestres, Paul Avillach, Gayo Diallo, Carlos Díaz Acedo, Johan van der Lei

First published: 4 December 2012 Full publication history

DOI: 10.1002/pds.3375 View/save citation

Abstract

PURPOSE:

Pharmacovigilance methods have advanced greatly during the last decades, making post-market drug assessment an essential drug evaluation component. These methods mainly rely on the use of spontaneous reporting systems and health information databases to collect expertise from huge amounts of real-world reports. The EU-ADR Web Platform was built to further facilitate accessing, monitoring and exploring these data, enabling an in-depth analysis of adverse drug reactions risks.

METHODS:

The EU-ADR Web Platform exploits the wealth of data collected within a large-scale European initiative, the EU-ADR project. Millions of electronic health records, provided by national health agencies, are mined for specific drug events, which are correlated with literature, protein and pathway data, resulting in a rich drug-event dataset. Next, advanced distributed computing methods are tailored to coordinate the execution of data-mining and statistical analysis tasks. This permits obtaining a ranked drug-event list, removing spurious entries and highlighting relationships with high risk potential.

RESULTS:

The EU-ADR Web Platform is an open workspace for the integrated analysis of pharmacovigilance datasets. Using this software, researchers can access a variety of tools provided by distinct partners in a single centralized environment. Besides performing standalone drug-event assessments, they can also control the pipeline for an improved batch analysis of custom datasets. Drug-event pairs can be substantiated and statistically analysed within the platform's innovative working environment.

CONCLUSIONS:

A pioneering workspace that helps in explaining the biological path of adverse drug reactions was developed within the EU-ADR project consortium. This tool, targeted at the pharmacovigilance community, is available online at https://bioinformatics.ua.pt/euadr/.

PLoS One. 2013; 8(12): e83016. PMCID: PMC3859628
Published online 2013 Dec 11. doi: 10.1371/journal.pone.0083016

Gathering and Exploring Scientific Knowledge in Pharmacovigilance

Pedro Lopes,[1] Tiago Nunes,[1] David Campos,[1] Laura Ines Furlong,[2] Anna Bauer-Mehren,[2] Ferran Sanz,[2] Maria Carmen Carrascosa,[2] Jordi Mestres,[2] Jan Kors,[3] Bharat Singh,[3] Erik van Mulligen,[3] Johan Van der Lei,[3] Gayo Diallo,[4] Paul Avillach,[4,5] Ernst Ahlberg,[6] Scott Boyer,[6] Carlos Diaz,[7] and José Luís Oliveira[1,*]

Dermot Cox, Editor

Abstract

Pharmacovigilance plays a key role in the healthcare domain through the assessment, monitoring and discovery of interactions amongst drugs and their effects in the human organism. However, technological advances in this field have been slowing down over the last decade due to miscellaneous legal, ethical and methodological constraints. Pharmaceutical companies started to realize that collaborative and integrative approaches boost current drug research and development processes. Hence, new strategies are required to connect researchers, datasets, biomedical knowledge and analysis algorithms, allowing them to fully exploit the true value behind state-of-the-art pharmacovigilance efforts. This manuscript introduces a new platform directed towards pharmacovigilance knowledge providers. This system, based on a service-oriented architecture, adopts a plugin-based approach to solve fundamental pharmacovigilance software challenges. With the wealth of collected clinical and pharmaceutical data, it is now possible to connect knowledge providers' analysis and exploration algorithms with real data. As a result, new strategies allow a faster identification of high-risk interactions between marketed drugs and adverse events, and enable the automated uncovering of scientific evidence behind them. With this architecture, the pharmacovigilance field has a new platform to coordinate large-scale drug evaluation efforts in a unique ecosystem, publicly available at http://bioinformatics.ua.pt/euadr/.

3.3.1 WORK DEVELOPED

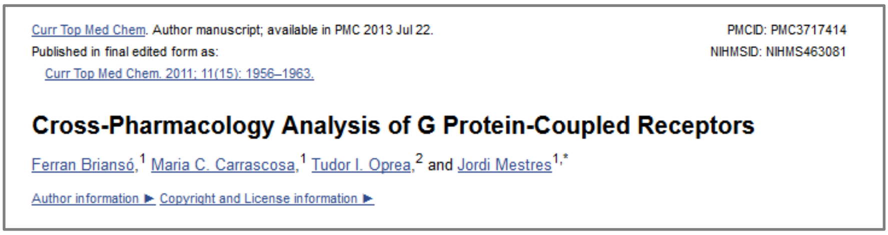
The EU-ADR was a research and development project funded by the Information and Communication Technologies (ITC) area of the European Commission under the VII Framework Programme that started in February 2008 and ended in January 2012. The main objective of the project was to develop an innovative computerised system to detect adverse drug reactions (ADRs), supplementing spontaneous reporting systems.

For this project, we developed a web service called cglAlertService. The web service was developed using Java, the Eclipse IDE, Apache Tomcat and Axis 2. We used a top-down approach that starts from the web service description, which is the WSDL, and then goes on to expose the web service.

The following table shows the name and the description of the two operations provided by the web service:

getSmileFromATC
This method accepts as input a drug encoded by the ATC (Anatomical Therapeutic Chemical) code at the 7-digits level and provides as output the chemical structure using SMILES (Simplified Molecular Input Line Entry Specification).
getUniprotListFromSmile
This method accepts as input a drug or metabolite encoded by a SMILES and returns a list of proteins that are related to the drug (Drug-Target Profile). We used known drug-target associations from public databases (AffinDB, BindingDB, ChemblDB, DrugBank, hGPCRlig, IUPHARdb, MOAD, Nracl, PDSP and PubChem) and extended them with *in silico* predictions [35]**. Drug metabolites were kindly provided by one of the partners (AstraZeneca) and were also processed with *in silico* target profiling methods** [36]**. Also provided is the evidence that supports each drug-target relationship, such as the binding affinity of the compound to the protein or the source database.**

3.4 BlastXP



Curr Top Med Chem. Author manuscript; available in PMC 2013 Jul 22.
Published in final edited form as:
Curr Top Med Chem. 2011; 11(15): 1956–1963.

PMCID: PMC3717414
NIHMSID: NIHMS463081

Cross-Pharmacology Analysis of G Protein-Coupled Receptors

Ferran Briansó,[1] Maria C. Carrascosa,[1] Tudor I. Oprea,[2] and Jordi Mestres[1,*]

Author information ► Copyright and License information ►

Abstract

The degree of applicability of chemogenomic approaches to protein families depends on the accuracy and completeness of pharmacological data and the corresponding level of pharmacological similarity observed among their protein members. The recent public domain availability of pharmacological data for thousands of small molecules on 204 G protein-coupled receptors (GPCRs) provides a firm basis for an in-depth cross-pharmacology analysis of this superfamily. The number of protein targets included in the cross-pharmacology profile of the different GPCRs changes significantly upon varying the ligand similarity and binding affinity criteria. However, with the exception of muscarinic receptors, aminergic GPCRs distinguish themselves from the rest of the members in the family by their remarkably high levels of pharmacological similarity among them. Clusters of non-GPCR targets related by cross-pharmacology with particular GPCRs are identified and the implications for unwanted side-effects, as well as for repurposing opportunities, discussed.

3.4.1 INTRODUCTION

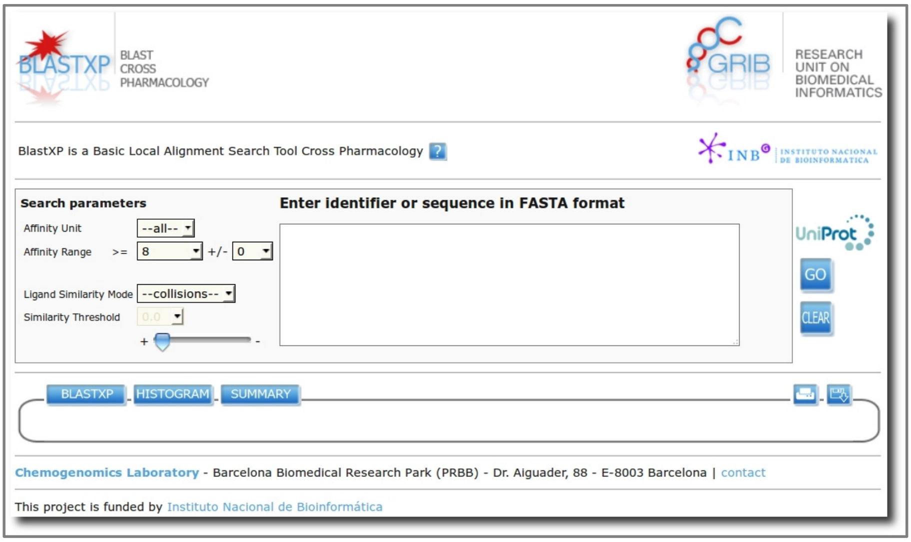
BlastXP was a web application developed in the GRIB-IMIM's Chemogenomics Lab to provide a graphical user interface to the methodology developed by the Lab to predict cross-pharmacology between targets. This methodology was thoroughly described and discussed in a publication in Curr. Top. Med. Chem. [33].

The level of cross-pharmacology between two target proteins is defined by the number of similar bioactive ligands that they share, which in turn reflect the probability that a small molecule being active on one target may also be active on the other. Thus, the cross-pharmacology profile of a given target protein is determined here by the list of other targets having at least one similar bioactive ligand in common with the given one. This number of similar shared bioactive ligands between two target proteins will be referred from now on as *cross-pharmacology score*, *raw score* or simply *score*.

Following the strategy of the Lab we developed the BlastXP platform that was made publicly available on the web. Unfortunately, due to technical reasons, the application is no longer available on the group's website. Plans to restore it in a near future are being contemplated. But it is important to mention it here since it represented our first step towards developing Pharmatrek, CT-link and ultimately CLARITY.

3.4.2 Application overview

Starting from a target protein, uniquely identified by its UniProt Accession Number or its FASTA sequence, BlastXP (Figure 10) provides the cross-pharmacology profile of that given target against all other targets in our database, and computed the expected probability of finding each cross-pharmacology relationship by chance.



BlastXP is a Basic Local Alignment Search Tool Cross Pharmacology ?

Search parameters

Affinity Unit --all--

Affinity Range >= 8 +/- 0

Ligand Similarity Mode --collisions--

Similarity Threshold 0.0

+ -

Enter identifier or sequence in FASTA format

Chemogenomics Laboratory - Barcelona Biomedical Research Park (PRBB) - Dr. Aiguader, 88 - E-8003 Barcelona | contact

This project is funded by Instituto Nacional de Bioinformática

Figure 10: BlastXP web interface.

3.4.2.1 Three-layer architecture

Implemented with a standard web three-layer architecture, BlastXP made use of an Apache Tomcat Application Server (apache-tomcat-6.0.18), Java (version 4.1.1), Javascript, Ajax and a MYSQL database (Server version: 5.0.51a). As shown in Figure 11, BlastXP used JSP as the presentation layer, Java servlet as the application layer, and MySQL as data managing layer.

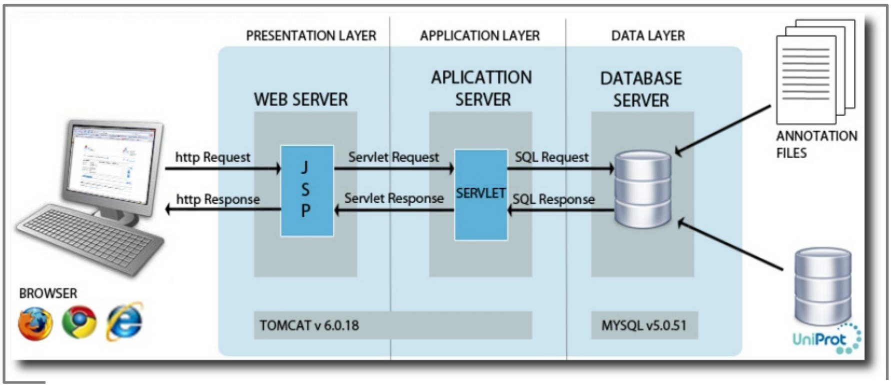


Figure 11: BlastXP three-layer architecture.

3.4.2.2 DATABASE SUMMARY

The MySQL database layer stored information about 4,263 targets (including Enzymes, GPCRs, Nuclear Receptors, and other Transporters and Channels), 245,709 ligands and the interactions between them. It included 725,240 target-ligand annotations, from which more than 60% had *pActivity (pK_i, pIC_{50} or pEC_{50})* values equal or above 6.0 (that means with interactions at or beyond the micromolar affinity level). The database was designed and implemented to provide a quick and efficient access to data. Figure 12 shows the data model.

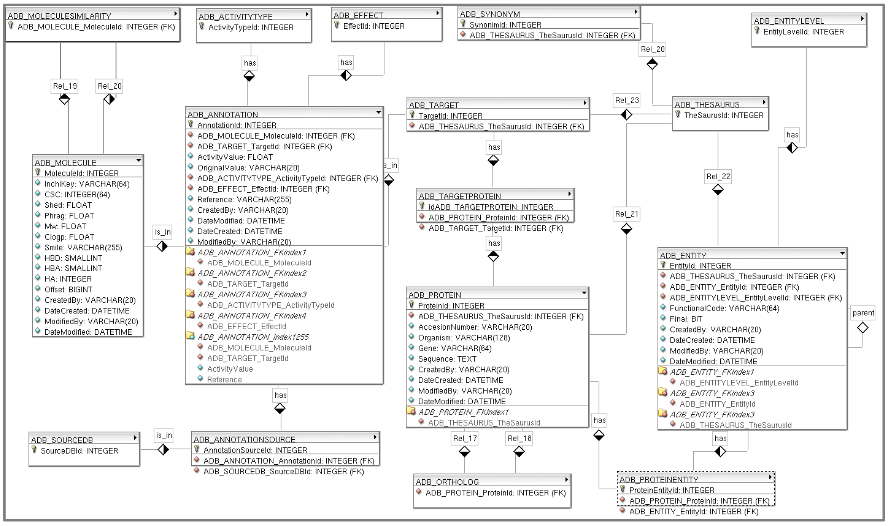
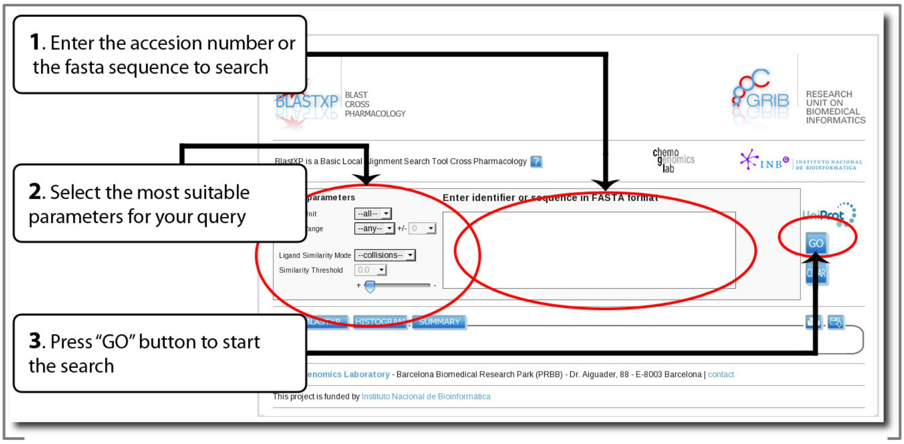


Figure 12: BlastXP database diagram.

3.4.2.3 BLASTXP RESULTS

To obtain the cross-pharmacology profile of a target protein with BlastXP, the user had to follow the three steps highlighted in Figure 13.



1. Enter the accesion number or the fasta sequence to search

2. Select the most suitable parameters for your query

3. Press "GO" button to start the search

BlastXP is a Basic Local Alignment Search Tool Cross Pharmacology

enomics Laboratory - Barcelona Biomedical Research Park (PRBB) - Dr. Aiguader, 88 - E-8003 Barcelona | contact

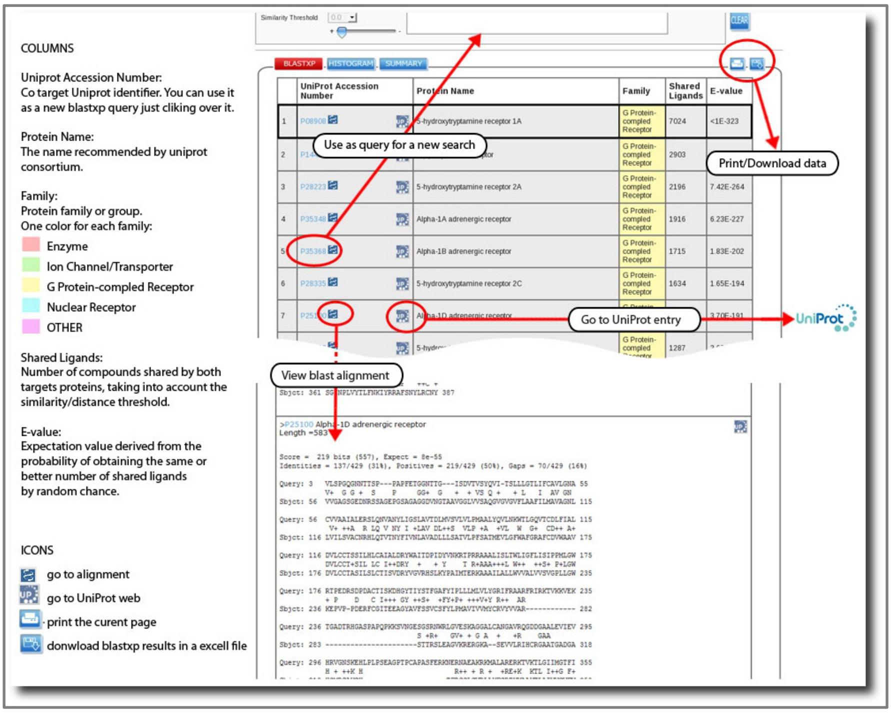
This project is funded by Instituto Nacional de Bioinformática

Figure 13: How to launch a query in BlastXP.

The application showed its results in three tabs, namely *BLASTXP*, *HISTOGRAM*, and *SUMMARY* (see Figure 14 to Figure 16).

3.4.2.3.1 BlastXP Tab

The BLASTXP tab (see Figure 14) showed the list with all the targets that had cross-pharmacology with the query protein. Results were sorted in a table from the most significant to the least significant cross-pharmacology E-value. Each row of the table showed the accession number of the protein with cross-pharmacology with the query target, its name, the family to



COLUMNS

Uniprot Accession Number:
Co target Uniprot identifier. You can use it as a new blastxp query just cliking over it.

Protein Name:
The name recommended by uniprot consortium.

Family:
Protein family or group.
One color for each family:

- 🟥 Enzyme
- 🟩 Ion Channel/Transporter
- 🟨 G Protein-compled Receptor
- 🟦 Nuclear Receptor
- 🟪 OTHER

Shared Ligands:
Number of compounds shared by both targets proteins, taking into account the similarity/distance threshold.

E-value:
Expectation value derived from the probability of obtaining the same or better number of shared ligands by random chance.

ICONS

- go to alignment
- go to UniProt web
- print the curent page
- donwload blastxp results in a excell file

Similarity Threshold 0.0

	UniProt Accession Number	Protein Name	Family	Shared Ligands	E-value
1	P08908	5-hydroxytryptamine receptor 1A	G Protein-compled Receptor	7024	<1E-323
2	P14	tor	G Protein-compled Receptor	2903	
3	P28223	5-hydroxytryptamine receptor 2A	G Protein-compled Receptor	2196	7.42E-264
4	P35348	Alpha-1A adrenergic receptor	G Protein-compled Receptor	1916	6.23E-227
5	P35368	Alpha-1B adrenergic receptor	G Protein-compled Receptor	1715	1.83E-202
6	P28335	5-hydroxytryptamine receptor 2C	G Protein-compled Receptor	1634	1.65E-194
7	P25100	Alpha-1D adrenergic receptor			3.70E-191
		5-hydr	G Protein-compled	1287	

```
Sbjct: 361 SGNPLVYTLFNKIYRRAFSNYLRCNY 387
```

```
>P25100 Alpha-1D adrenergic receptor
Length =583

Score =  219 bits (557), Expect = 8e-55
Identities = 137/429 (31%), Positives = 219/429 (50%), Gaps = 70/429 (16%)

Query: 3   VLSPGQGNNTTSP---PAPFETGGNTTG---ISDVTVSYQVI-TSLLLGTLIFCAVLGNA 55
            V+  G G +  S    P     GG+  G    +  + VS Q +   + L   I  AV GN
Sbjct: 56  VVGAGSGEDNRSSAGEPGSAGAGGDVNGTAAVGGLVVSAQGVGVGVFLAAFILMAVAGNL 115

Query: 56  CVVAAIALERSLQNVANYLIGSLAVTDLMVSVLVLPMAALYQVLNKWTLGQVTCDLFIAL 115
             V+ ++A  R LQ V NY I +LAV DL++S  VLP +A  +VL  W  G+  CD++ A+
Sbjct: 116 LVILSVACNRHLQTVTNYFIVNLAVADLLLSATVLPFSATMEVLGFWAFGRAFCDVWAAV 175

Query: 116 DVLCCTSSILHLCAIALDRYWAITDPIDYVNKRTPRRAAALISLTWLIGFLISIPPMLGW 175
            DVLCCT+SIL LC I++DRY  +   + Y    T R+AAA+++L W++  ++S+ P+LGW
Sbjct: 176 DVLCCTASILSLCTISVDRYVGVRHSLKYPAIMTERKAAAILALLWVVALVVSVGPLLGW 235

Query: 176 RTPEDRSDPDACTISKDHGYTIYSTFGAFYIPLLLMLVLYGRIFRAARFRIRKTVKKVEK 235
            + P    D   C I+++ GY ++S+  +FY+P+ +++V+Y R++  AR
Sbjct: 236 KEPVP-PDERFCGITEEAGYAVFSSVCSFYLPMAVIVVMYCRVYVVAR------------ 282

Query: 236 TGADTRHGASPAPQPKKSVNGESGSRNWRLGVESKAGGALCANGAVRQGDDGAALEVIEV 295
                                      S +R+   GV+ + G A  +   +R    GAA
Sbjct: 283 ----------------------STTRSLEAGVKRERGKA--SEVVLRIHCRGAATGADGA 318

Query: 296 HRVGNSKEHLPLPSEAGPTPCAPASFERKNERNAEAKRKMALARERKTVKTLGIIMGTFI 355
            H + ++K H                         R++ + R +  +RE+K  KTL I++G F+
```

Figure 14: BlastXP help window explaining what is showwn in BLASTXP results tab.

which belongs, the number of ligands that shares with the query target and the calculated cross-pharmacology E-value.

Proteins presenting a Blast similarity (by sequence identity) with the query target had its background color highlighted in gray. The *Family* column was also coloured in order to easily identify the protein family to which the targets belonged: orange for enzymes, green for ion channels and transporters, yellow for GPCRs, blue for nuclear receptors and pink for others. Of course, the query target should always appear having cross-pharmacology with itself, so the row for the query target in the results was marked with a black bolded frame.

All proteins that had cross-pharmacology with the query target were shown even if the E-values were not relevant. In this case, those rows with an E-value greater than 1 were depicted in a soft gray style. The user could use any target from the results list to perform a new search just by clicking on its accession number. All accession numbers also showed an icon to link to UniProt. Additionally, all entries having a significant Blast alignment with the query protein showed an icon as a shortcut to access the alignment below.

3.4.2.3.2 Histogram Tab

The HISTOGRAM tab represented partially the cross-pharmacology profile of the given target. It showed the number of similar shared ligands with every protein, including only those proteins with E-value lower than 1, and sorting the profile by ascending E-value. The bars in the histrogram were colored according to the protein family. In Figure 15 we can see an example with the *Alpha-2A adrenergic receptor* (P08913) as target query.

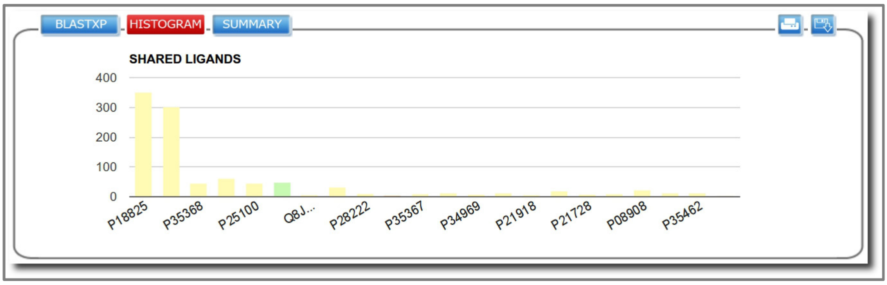
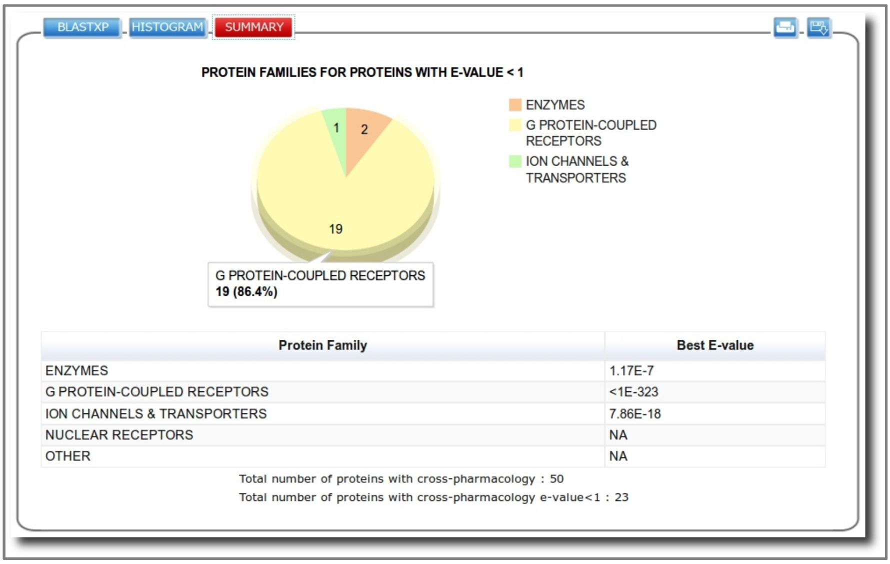


Figure 15: BlastXP results Histogram tab.

3.4.2.3.3 Summary Tab

Finally, the SUMMARY tab collected the results grouped by families. As shown in Figure 16, we used the JS Charts Javascript library to generate a pie chart with the proportion of related targets belonging to each main protein family. A table with a list of the best E-values found in each group was also presented, as well as the total count of proteins with at least one similar shared ligand (cross-pharmacology score higher than 0) and the number of these proteins with cross-pharmacology E-value below 1.0.


BLASTXP
HISTOGRAM
SUMMARY
PROTEIN FAMILIES FOR PROTEINS WITH E-VALUE < 1
1
2
19
ENZYMES
G PROTEIN-COUPLED RECEPTORS
ION CHANNELS & TRANSPORTERS
G PROTEIN-COUPLED RECEPTORS
19 (86.4%)

Protein Family	Best E-value
ENZYMES	1.17E-7
G PROTEIN-COUPLED RECEPTORS	<1E-323
ION CHANNELS & TRANSPORTERS	7.86E-18
NUCLEAR RECEPTORS	NA
OTHER	NA

Total number of proteins with cross-pharmacology : 50
Total number of proteins with cross-pharmacology e-value<1 : 23

Figure 16: Results summary tab.

As we can see, the application was a first attempt to provide a user-friendly interface to the user to execute a complex software (that computes XP) easily. Moreover, we provided an interactive graphical interface to explore the results using simple charts (pie charts and bar charts) and revealed new insights in the data.

3.5 Paper Pharmatrek

Carrascosa MC, Massaguer OL, Mestres J. PharmaTrek: A Semantic Web Explorer for Open Innovation in Multitarget Drug Discovery. Mol Inform. 2012 Aug;31(8):537–41. DOI: 10.1002/minf.201200070

3.5.1 TECHNICAL RESULTS

Being a Communication article, the Pharmatrek paper doesn’t include information about the technical details of the platform. In the following sections, more detailed information about the architecture, functionality and implementation of the software will be provided.

3.5.1.1 FUNCTIONALITY

To explore the data contained in the OPS platform, the Pharmatrek application offers these functionalities:

- Search targets by name and show information available for these targets.
- Select multiple targets of interest.
- Apply pharmacology filters such activity type or value range for each selected target.
- Show pharmacology information of the selected targets as an interactive heatmap, compound *vs* target
- Given a heatmap, expand it adding to the visualization other targets that also have pharmacology interaction with the obtained ligands.
- Be able to add new targets of interest according to the results obtained.

- Download results as CSV text files.
- Search ligands by name and by structure
- Select multiple ligands of interest and show their pharmacological information as an interactive heatmap and be able to apply pharmacological filters during the process.
- Connection with other Open PHACTS example applications, in particular with the ChemBioNavigator. The ChemBioNavigator allows the user to visualize the chemical and biological space of a molecule group in a chemical-aware manner.

The application evolved during the project. In the last version, an important new functionality was added:

- Show information about the pharmacological predictions generated by the PredictFX software [35] developed by Chemotargets.

3.5.1.2 DESIGN

The application is designed under a client-server architecture with a presentation layer developed using Flex, an application layer developed using Scala, and a data layer represented by the OPS platform.

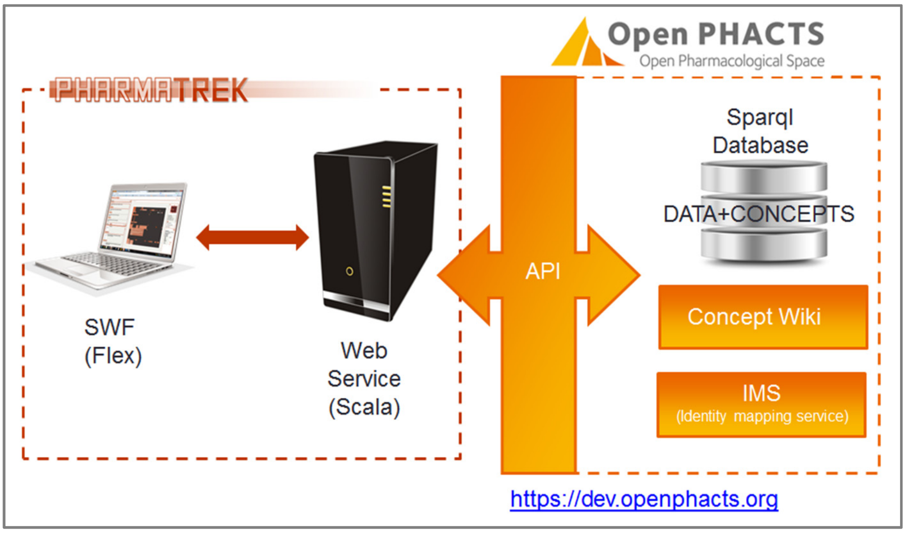


Figure 17: Pharmatrek architecture.

In the last version of the application, the data layer was modified to add access to the pharmacological predictions generated by the PredictFX software.

3.5.1.3 DATA SOURCES

Open PHACTS (Open Pharmacological Concept Triple Store) is a European project that started in 2011 funded by the Innovative Medicines Initiative (IMI). The goal of the project was to deliver and maintain an 'open pharmacological space' (OPS) integrating dispersed and heterogeneous data sources and provide also powerful tools to enable scientists to explore this unified data repository to extract knowledge to improve the process of drug discovery.

To perform this goal, they use semantic web standards and technologies. According to the World Wide Web Consortium

(W3C), “the semantic web provides a common framework that allows data to be shared and reused across application, enterprise, and community boundaries” and goes even further by stating that it is “a web of data that can be processed directly and indirectly by machines”. This is one of the main reasons why Open PHACTS adopted semantic web tools to create a data integration platform that was going to be made available in a manner that facilitates computational processing. To create this data platform, the project converted data from relevant sources (PubChem, DrugBank, ChemSpider) into RDF (resource description framework), a standard model for data interchange on the web. With the RDF format, data is stored as triples, in the form subject-predicate-object.

To access this data platform, they also provided a rich Application Programming Interface (API). The API provides the developer an standard web API [18], the API is, in turn, developed using SPARQL [37], a semantic query language to retrieve and manipulate data stored in RDF datasets [38][39]. The API encapsulates the access to the data providing a simpler way to interrogate the data than using SPARQL directly.

Pharmatrek is an example of the applications developed in the project that use this API to access the data.

In the following table, the API calls used to implement the application functionality are mentioned with a short description:

API call	Description
/search/freetext	Return a list of conceptWIKi URIs for a user specified free text search term.
/target	Return information about a single target.
/target/pharmacology/pages	Return pharmacological data for a user specified target with filtering options (target organism, activity type, activity value, etc.)
/compound	Return information about a single compound
/compound/pharmacology/pages	Return pharmacological data for a user specified compound with filtering options (target organism, activity type, activity value, etc.)

/structure/exact	Return a ChemSpider URI corresponding to the input SMILE string.
/structure/similarity	A list of ChemSpider URIs for compounds similar to the input molecule will be returned
/structure/substructure	A list of ChemSpider URIs for compounds containing the specified structure will be returned.

In the last part of the project, we established a collaboration with Chemotargets. We used PredictFX, a systems drug discovery package developed by Chemotargets. PredictFX is designed as a modular computational framework that links together different input and output spaces through appropriate connections. In particular, we implemented the connections between ligands and targets. We used version 15 of the ChEMBL database as the input of the PredictFX software. This version of ChEMBL contains 1.194.038 compounds. With PredictFX, we enriched the OPS data with 69.205.000 of pharmacology predictions between 1.089.490 compounds and 4.496 targets defined in the OPS platform (Figure 18).

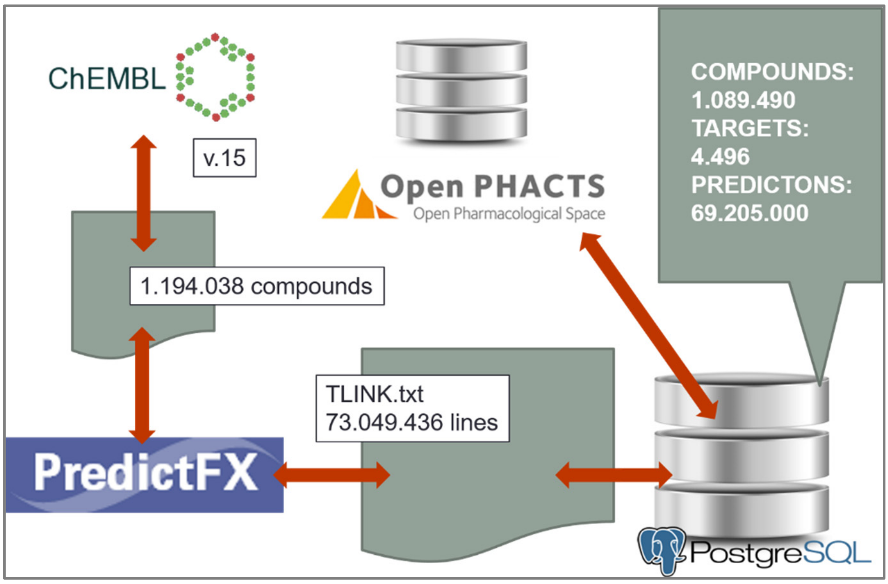
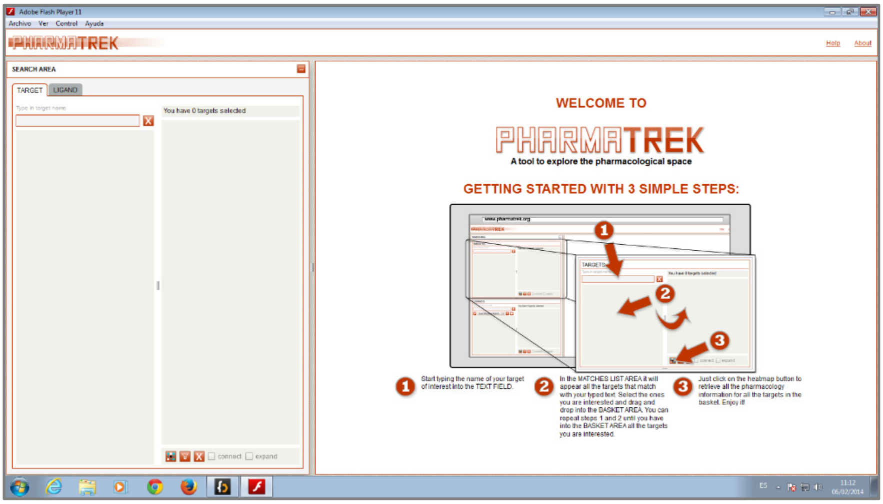


Figure 18: Pharmatrek and PredictFX.

3.5.1.4 GUI AND USE CASE

The application was located at http://www.pharmatrek.org and it was open to the public domain. But due to the lack of backforward compatibility of the API provided by the Open PHACTS project, and the lack of resources for support and maintenance, Pharmatrek stopped working after the end of our participation in the project. Figure 19 shows a snapshot of its main page. As we can see, it is different from the picture published in the paper. This is due to the fact that the application evolved during the years of the project. In the figure, we can see a *Search area* where the user could select “Target” or “Ligand” and a *Help area* with a quick help to start using the application following just 3 steps. As an application example, we will show how to address the question of “extracting all potent Factor Xa inhibitors with selectivity of at least two orders of magnitude relative to thrombin and trypsin, two phylogenetically related serine proteases”.



You have 0 targets selected

WELCOME TO

PHARMATREK

A tool to explore the pharmacological space

GETTING STARTED WITH 3 SIMPLE STEPS:

1 Start typing the name of your target of interest into the TEXT FIELD.

2 In the MATCHES LIST AREA it will appear all the targets that match with your typed text. Select the ones you are interested and drag and drop into the BASKET AREA. You can repeat steps 1 and 2 until you have into the BASKET AREA all the targets you are interested.

3 Just click on the heatmap button to retrieve all the pharmacology information for all the targets in the basket. Enjoy it!

Figure 19: Pharmatrek main page.

3.5.1.4.1 *Defining the target profile*

To start with, the user has to enter the target names into the field located in the middle of the page. It is a Google-like search field that suggests names of targets while you are entering your search. The idea is that this field helps you to find the target that you are interested in.

You are then automatically redirected to two tabs:

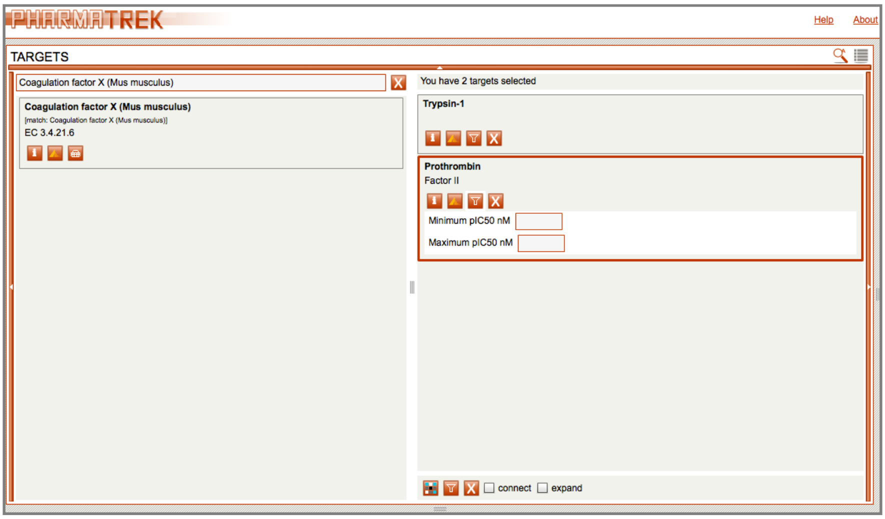
- The *Search tab* that includes the search area and the basket area

- The *Browser tab* with all the targets that are in the heatmap.

For example, if the user enters “coagulation Factor X (Mus musculus)”, it will match with the enzyme EC 3.4.21.6. The user will have the possibility to get more details about the enzyme by clicking on the icon “Show details” and “View this target in Explorer” or “Add to the basket”. For the last option, the user will then be able to create a heatmap in order to visualize all the compounds that show some bioactivities (such as K_i, K_d, IC_{50} or EC_{50}) for this enzyme.

3.5.1.4.2 The basket selection

An interesting feature of the *Basket selection* is the possibility to integrate several targets in the query and to filter the activity of interest for each target. For example, the user can address a complex query such as retrieve all the ligands having a –log(Activity) value larger than or equal to 7.5 for the coagulation factor Xa and being at least two orders of magnitude selective against trypsin and thrombin (two phylogenetically related serine proteases). With the “Connect checkbox”, the heatmap is restricted to only the ligands that have some activity against all the targets in the basket whereas the “Expand checkbox” will expand the heatmap by adding targets that have some activity against the heatmap ligands.



Coagulation factor X (Mus musculus)

Coagulation factor X (Mus musculus)
[match: Coagulation factor X (Mus musculus)]
EC 3.4.21.6

You have 2 targets selected

Trypsin-1

Prothrombin
Factor II

Minimum pIC50 nM
Maximum pIC50 nM

□ connect □ expand

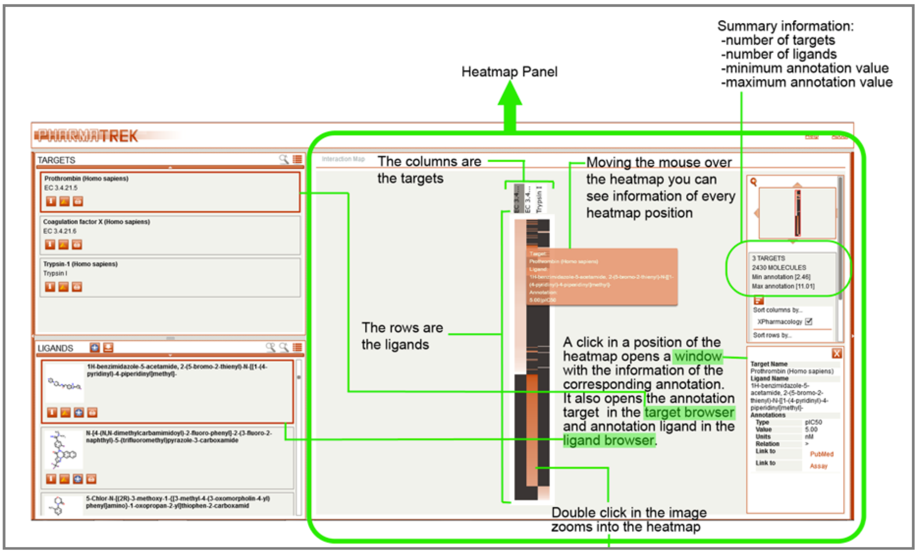
Figure 20: Target search layout.

3.5.1.4.3 The interactive heatmap

Once you have entered the target or target profile and all the filters that you want to apply, you can click on the “Show heatmap” button to obtain an interactive map containing all protein-ligand interactions that meet all criteria defined in the basket. As can be observed in the picture added below, an interactive heatmap appears in the Heatmap Panel.

In the interaction map, rows represent ligands and columns the targets defined in the basket.

To keep only those ligands that have interaction with all the targets in your search, you can click on the “connect” check box located at the bottom of the basket field. If you pass the mouse over the heatmap, a tooltip appears with the value of the interaction (in –logarithmic scale) of the corresponding ligand/row-target/column.



Summary information:
-number of targets
-number of ligands
-minimum annotation value
-maximum annotation value

Heatmap Panel

The columns are the targets

Moving the mouse over the heatmap you can see information of every heatmap position

The rows are the ligands

A click in a position of the heatmap opens a window with the information of the corresponding annotation. It also opens the annotation target in the target browser and annotation ligand in the ligand browser.

Double click in the image zooms into the heatmap

PHARMATREK

TARGETS

Prothrombin (Homo sapiens)
EC 3.4.21.5

Coagulation factor X (Homo sapiens)
EC 3.4.21.6

Trypsin-1 (Homo sapiens)
Trypsin I

LIGANDS

1H-benzimidazole-5-acetamide, 2-(5-bromo-2-thienyl)-N-[[1-(4-pyridinyl)-4-piperidinyl]methyl]-

N-[4-(N,N-dimethylcarbamimidoyl)-2-fluoro-phenyl]-2-(3-fluoro-2-naphthyl)-5-(trifluoromethyl)pyrazole-3-carboxamide

5-Chlor-N-[(2R)-3-methoxy-1-([3-methyl-4-(3-oxomorpholin-4-yl) phenyl]amino)-1-oxopropan-2-yl]thiophen-2-carboxamid

3 TARGETS
2430 MOLECULES
Min annotation [2.46]
Max annotation [11.01]

Sort columns by...
XPharmacology ☑
Sort rows by...

Target Name
Prothrombin (Homo sapiens)
Ligand Name
1H-benzimidazole-5-acetamide, 2-(5-bromo-2-thienyl)-N-[[1-(4-pyridinyl)-4-piperidinyl]methyl]-
Annotations

Type	pIC50
Value	5.00
Units	nM
Relation	>
Link to	PubMed
Link to	Assay

Figure 21: Heatmap layout.

At the right-hand side of the heatmap, you can see a summary with the number of targets and molecules displayed in the heatmap and also the maximum and minimum affinity values. There is also an image that represents the heatmap and a slider that allows you to make zoom in, zoom out, and move into the heatmap. The default color is from white, for a weak activity (Minimum annotation value), to red, for a strong activity annotated to the protein (maximum annotation value). It is grey when no annotation has been found. For the sake of convenience, one can also customise the colours of the heatmap with the colour selectors that are located at the bottom-right corner of the heatmap.

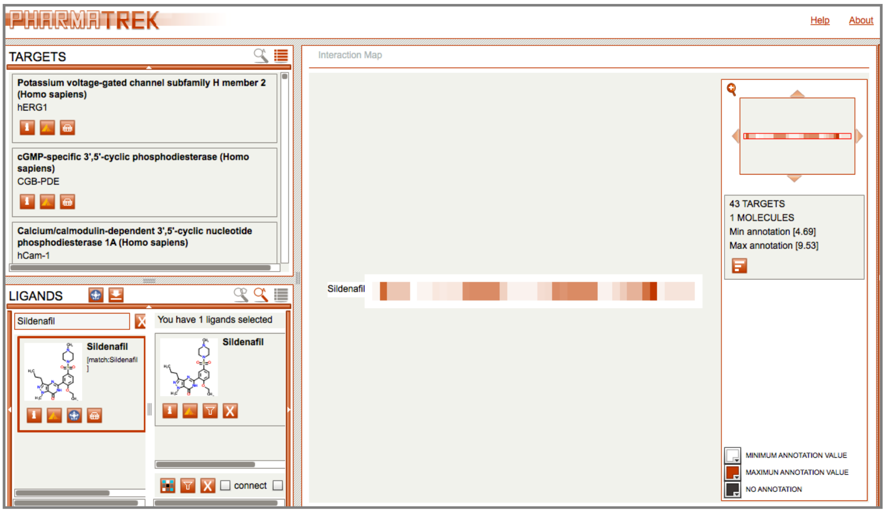
Last but not least, one can also ask Pharmatrek to check whether the compounds fulfilling the criteria for the target profile defined in the basket have additional interactions on other targets in the source(s) being explored. To do that, you can click in the “Expand target space” check box. Upon activation of this check box, the application makes a request to expand the target space with additional targets not included originally in the target profile defined in the basket. A useful feature is that any new target appearing after this expansion can be added to the basket. Simply put the mouse on top of the target name appearing in the labels of the columns (zoom in if necessary) and drag & drop the name into the basket. You can then apply new filters and perform a new ligand extraction request.

3.5.1.4.4 The ligand tab

At any point during the Pharmatrek session, the structures of all ligands fulfilling all criteria defined appear in the “Ligands” section located below the basket. By clicking the image of a structure, more detailed information about the ligand can be obtained, such as name, molecular formula, molecular weight, InChI, InChI-keys, SMILES, and any additional properties and descriptors that might be available in the source being explored. One can always save the structures in SMILES of all the ligands shown by clicking in the “Save smiles” button. An interesting feature is the connection to the ChemBioNavigator eApp. So, any ligand reported in the Pharmatrek session can then be transferred into ChemBioNavigator for further analysis of the ligand.

3.5.1.4.5 Defining the ligand profile

Finally, all the features described previously for the “target profile” are also applicable to the “ligand profile”, meaning that Pharmatrek will provide all the bioactivities annotated for a specific compound of interest. For example, a search for “sildenafil” will retrieve activities for 43 targets (see Figure 21 below).



Potassium voltage-gated channel subfamily H member 2 (Homo sapiens)
hERG1

cGMP-specific 3',5'-cyclic phosphodiesterase (Homo sapiens)
CGB-PDE

Calcium/calmodulin-dependent 3',5'-cyclic nucleotide phosphodiesterase 1A (Homo sapiens)
hCam-1

Sildenafil

Sildenafil
[match:Sildenafil
]

You have 1 ligands selected

Sildenafil

□ connect □

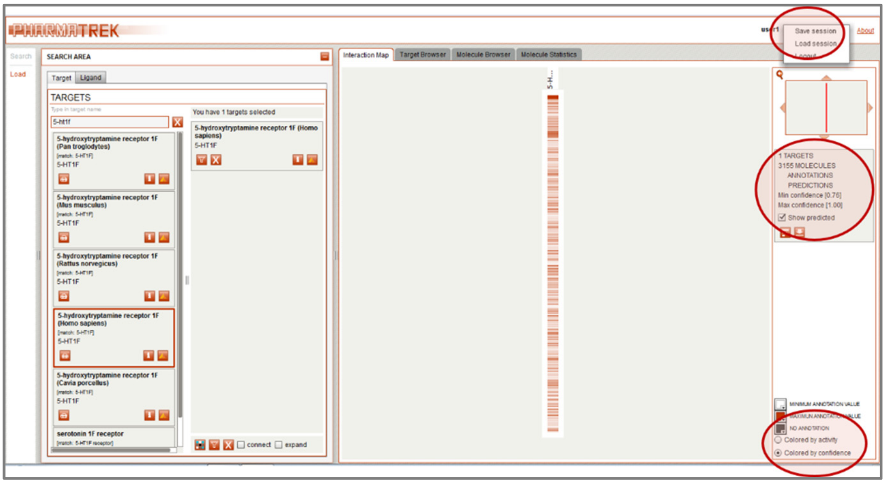
Sildenafil

43 TARGETS
1 MOLECULES
Min annotation [4.69]
Max annotation [9.53]

MINIMUM ANNOTATION VALUE
MAXIMUN ANNOTATION VALUE
NO ANNOTATION

Figure 22: Heatmap layout for sildenafil.

As mentioned in the previous section, in the last version of Pharmatrek, predicted pharmacological data was added to the application. As shown in the image below, the heatmap also displays information about the predictions generated by the PredictFX software.



5-ht1f

5-hydroxytryptamine receptor 1F (Pan troglodytes)
[match: 5-HT1F]
5-HT1F

5-hydroxytryptamine receptor 1F (Mus musculus)
[match: 5-HT1F]
5-HT1F

5-hydroxytryptamine receptor 1F (Rattus norvegicus)
[match: 5-HT1F]
5-HT1F

5-hydroxytryptamine receptor 1F (Homo sapiens)
[match: 5-HT1F]
5-HT1F

5-hydroxytryptamine receptor 1F (Cavia porcellus)
[match: 5-HT1F]
5-HT1F

serotonin 1F receptor
[match: 5-HT1F receptor]

You have 1 targets selected

5-hydroxytryptamine receptor 1F (Homo sapiens)
5-HT1F

□ connect □ expand

1 TARGETS
3155 MOLECULES
ANNOTATIONS
PREDICTIONS
Min confidence [0.76]
Max confidence [1.00]
☑ Show predicted

○ Colored by activity
● Colored by confidence

Figure 23: sildenafil predictions and known data.

A video with a demo of the Pharmatrek application is available here:

https://www.youtube.com/watch?v=nXLg8VXLREk

3.6 Paper CT-link (to be submitted)

CT-link: linking chemistry, pharmacology, and toxicology for early identification of safety issues in drug discovery

Maria C. Carrascosa,[a],[b] Nikita Remez,[b] Ricard Garcia-Serna,[b] and
Jordi Mestres*[a],[b]

Abstract: There is no doubt that the use of computational methods in drug discovery contributes to reducing the cost of the process. Among all computational methods used, predictive methods are of great importance. In order to improve the performance of these predictive methods, the amount and coverage of the data used is of critical importance. In recent years, the amount of the data available in biomedical research has expanded significantly. However, the availability of data is not sufficient. As new challenges appear, we need better software tools to exploit the data.

In particular, better tools to explore and visualize data are necessary to allow the user to extract new knowledge and insights from it. With this goal in mind, we developed CT-link, an advanced web-based graphical application that improves the exploration and navigation of data relevant to drug discovery. In particular, focus was given to design a new framework to link heterogeneous data coming from chemistry, pharmacology, and toxicology. The new platform provides an advanced graphical user interface to facilitate data analysis and knowledge extraction from it.

[a] *M. C. Carrascosa, J. Mestres Systems Pharmacology, GRIB*

[b] *N. Remez, R. Garcia-Serna, J. Mestres Chemotargets S.L.*

The drug discovery process is a costly process[1][2][3][4]. To reduce this cost, computational methods are used along the drug discovery pipeline. These computational methods use data to build predictive models. The amount of data available is increasing exponentially and it is having a tremendous impact on the quality of the results obtained. For example the current version of ChEMBL[5] contains more than two million compounds and over 14 million of pharmacological interactions. The last release of PubChem[6][7] contains also more than two million compounds with biological activities and more than one million bioassays. If beyond linking molecules and proteins with pharmacological data, one starts considering other aspects relevant to drug discovery, such as adverse reactions, levels of gene expression, and therapeutic area, and their corresponding associations, the computational methods used enter the realm of the big data paradigm[8][9].

The term Big Data is commonly associated with the three Vs that define properties or dimensions, namely, Volume, Variety and Velocity[10]. Volume refers to the magnitude of data, quite often measured in terabytes and petabytes; variety refers to the structural heterogeneity in a dataset; and velocity refers to the rate at which data are generated and the speed at which it should be analysed and acted upon. In the area of biomedical research, we specially deal with volume and variety. There is an increasing number of initiatives to exploit data following this paradigm in the area of biomedical research[11][12][13].

Related to the variety of data involved in the process of drug discovery, there is also an increasing need to find connections between the different types of data. For example, beyond exploring the links between molecules and proteins, one may be interested in establishing links between proteins and pathways, pathways and organs, and so on and so for. We can see the data to analyse as a large complex interconnected network. Here, we will focus on exploiting this complex network to analyse the various links between pairs of heterogeneous data and visualize them to extract new knowledge on the pharmacology and safety of molecules.

With this goal in mind, Chemotargets developed a predictive system called PredictFX [14]. It integrates a wide range of methodologies that, under a consensus approach, are used to predict the affinity profile of small molecules across thousands of proteins and its potential association with adverse reactions.

The software is a Linux command line tool developed in Python that generates as an output a high volume of plain text files. To explore these data files, the user has to import these files into third party platforms to visualize and analyse the data. This is not the optimal way of working because it forces the user to make a round trip between the predictive system and the analysis and visualization of data. Moreover, the scientific method is an interactive process of question formulation and response generation that, when separating the predictive system from the exploration and visualization system, the process is not fluent. In addition, to execute the PredictFX software, one makes often the assumption that the user has a certain level of computational knowledge to perform the various tasks, that is not always the case.

Therefore, even though PredictFX was proven to be a very powerful approach for the generation of pharmacology and safety predictions, the lack of a visualization component was a difficult barrier for the non-expert user.

Although data exploration and analysis has been used for a long time in scientific research[15], particularly since the advent of computer graphical user interfaces(GUI)[16][17], there is a growing need to develop and improve the tools and methods for interactive data analysis and visualization[18].

For this reason, we developed a new graphical user interface for the PredictFX software that allows the user to execute the various commands in a friendly manner and visualize the data in an intuitive and interactive way. The resulting platform is called CT-link.

In the CTlink platform, two main aspects were further developed. On one hand, in PredictFX the user could execute the software to obtain known and predicted bioactivities (T-link) and also to generate safety predictions (S-link). In CT-link, the user can also generate metabolites (M-link) and predict their bioactivies to be compared with those of the original substrate molecule (MT-link). Figure 1 summarizes the data flow of the platform.

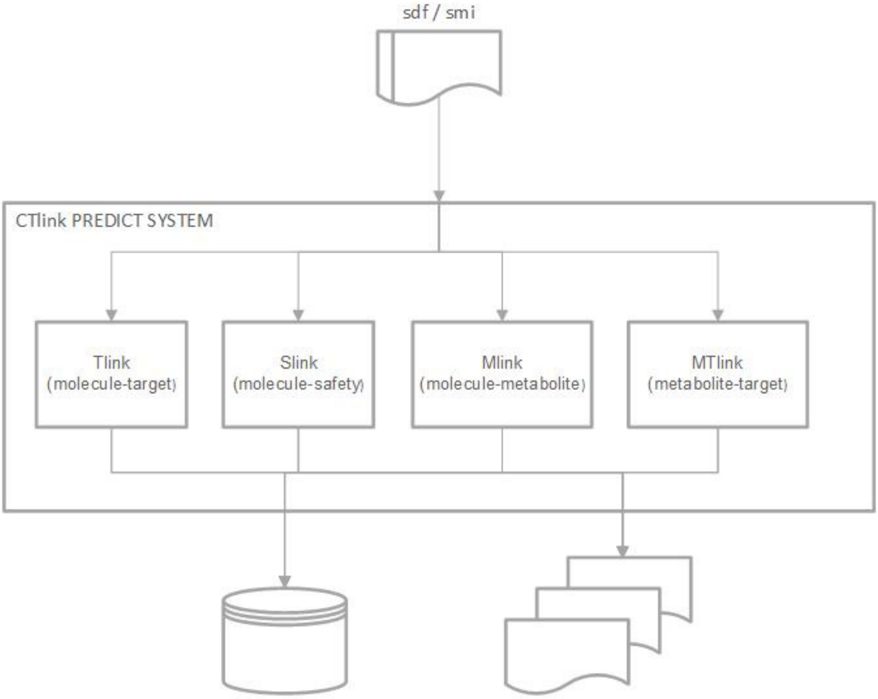


Figure 1: CT-link modular data flow diagram.

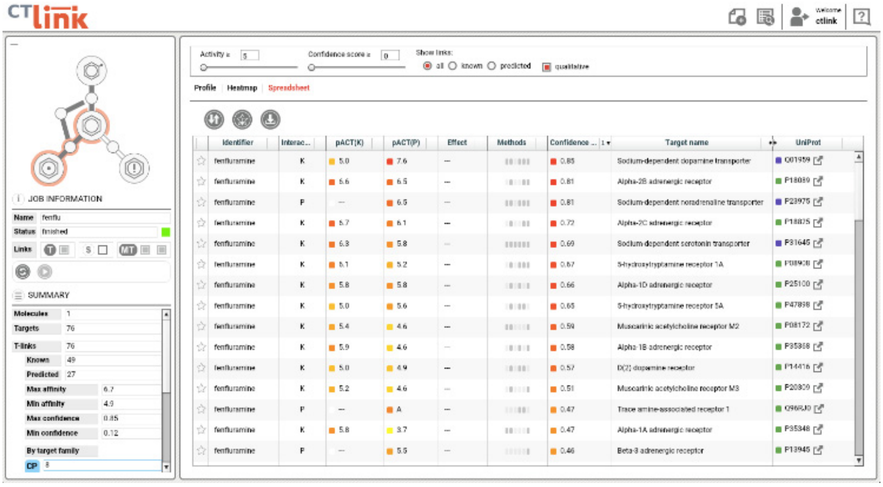
On the other hand, a web application was developed to execute and visualize all data generated. The development of this graphical user interface (GUI) also includes the design and implementation of a new architecture that connects the predictive system data generator and the data exploration system and provides advanced user-friendly exploration capabilities.

Access to the application by the user is done with a web browser. The user can upload a set of molecules of interest and with just two clicks get easily and quickly the predictions and known data for the input molecules. Molecules can be entered in the system in two different ways: by uploading the structures in mol files or SMILES formats, or by drawing the structures with a molecular sketcher. The system allows the user to edit the structures of all molecules uploaded.

Once the user has finished preparing the set of molecules, it can submit an execution to obtain different types of predictions: protein bioactivities, safety events and metabolite structures. To do that, the user has to select the type of predictions that wants to perform and simply click the “run” button. Once the job has started, the system shows the user the status of the execution, as well as an estimation of the remaining time. Results are delivered to the system as soon as they are produced so, at any point in time, the user can explore the results being obtained.

To explore the results, a graphical navigation network is shown. The user can navigate interactively through four entities, namely, molecules, proteins, safety and metabolites, and four links between them, namely, molecule-protein, molecule-safety, molecule-metabolite and metabolite-protein. For each entity, the user can see a summary of the results. For example, for each molecule the number of interacting targets, safety terms associated and metabolites generated is provided.

In the case of protein targets and safety terms, a bar chart showing the distribution of target families and safety categories is also displayed. In addition, regarding the links between the different entities, three options are available. The data is shown as a distribution bar chart, as a heatmap and as a spreadsheet. For example, for the molecule-protein links a stacked bar chart is shown with information on the protein families of therapeutic relevance. Also, a heatmap is produced to allow the used to explore the molecule-protein interactions in more detail. Finally, the data can be displayed as a classical spreadsheet (Figure 2). The application also provides substantiation information that is activated by clicking on any row of the spreadsheet. At any point of the application, the user can select, filter and sort data and this is applied to all links. Finally, the application also provides download and reporting tools to save the results obtained.



JOB INFORMATION

Name	fenflu
Status	finished
Links	T □ S □ MT

SUMMARY

Molecules	1
Targets	76
T-links	76
Known	49
Predicted	27
Max affinity	6.7
Min affinity	4.9
Max confidence	0.85
Min confidence	0.12
By target family	
CP	8

Activity ≥ 5 Confidence score ≥ 0 Show links: ● all ○ known ○ predicted ☒ qualitative

	Identifier	Interac...	pACT(K)	pACT(P)	Effect	Methods	Confidence ... 1 ▾	Target name	UniProt
☆	fenfluramine	K	5.0	7.6	—		0.85	Sodium-dependent dopamine transporter	Q01959
☆	fenfluramine	K	6.6	6.5	—		0.81	Alpha-2B adrenergic receptor	P18089
☆	fenfluramine	P	—	6.5	—		0.81	Sodium-dependent noradrenaline transporter	P23975
☆	fenfluramine	K	5.7	6.1	—		0.72	Alpha-2C adrenergic receptor	P18825
☆	fenfluramine	K	6.3	5.8	—		0.69	Sodium dependent serotonin transporter	P31645
☆	fenfluramine	K	6.1	5.2	—		0.67	5-hydroxytryptamine receptor 1A	P08908
☆	fenfluramine	K	5.8	5.8	—		0.66	Alpha-1D adrenergic receptor	P25100
☆	fenfluramine	K	5.0	5.6	—		0.65	5-hydroxytryptamine receptor 5A	P47898
☆	fenfluramine	K	5.4	4.6	—		0.59	Muscarinic acetylcholine receptor M2	P08172
☆	fenfluramine	K	5.9	4.6	—		0.58	Alpha-1B adrenergic receptor	P35368
☆	fenfluramine	K	5.0	4.9	—		0.57	D(2) dopamine receptor	P14416
☆	fenfluramine	K	5.2	4.6	—		0.51	Muscarinic acetylcholine receptor M3	P20309
☆	fenfluramine	P	—	A	—		0.47	Trace amine-associated receptor 1	Q96RJ0
☆	fenfluramine	K	5.8	3.7	—		0.47	Alpha-1A adrenergic receptor	P35348
☆	fenfluramine	P	—	5.5	—		0.46	Beta-3 adrenergic receptor	P13945

Figure 2: Fenfluramine target profile spreadsheet.

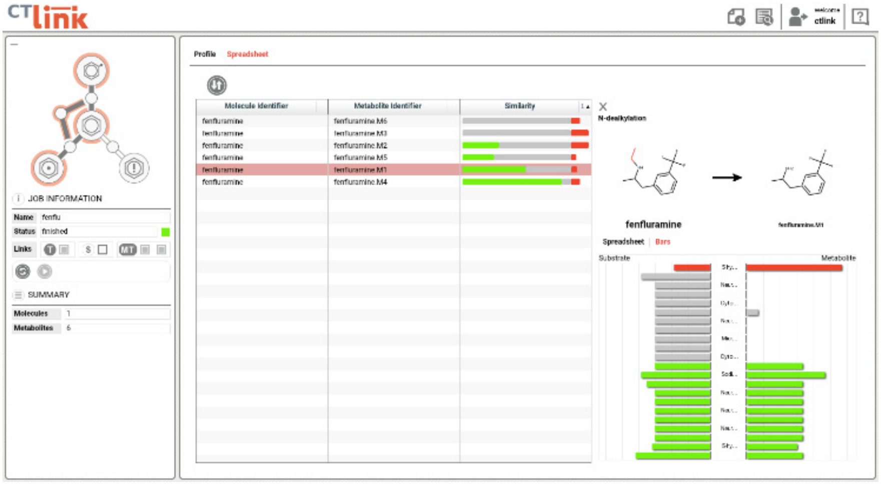
As an example of the use of the application, Figure 2 shows the pharmacology results for fenfluramine. Fenfluramine is an anorectic that was introduced in the US market in 1973 and withdrawn in 1997 after reports of heart valve disease. Fenfluramine represents a particularly interesting case since its cardiotoxicity effects are associated with the formation of a metabolite, norfenfluramine, that acts as a potent 5-HT$_{2B}$ agonist, leading to inappropriate valve cell division and ultimately causing damage to the heart valves that continue long after stopping the medication. Using CT-link, we are able to address this particular case due to the capacity of the platform to integrate predictive off-target pharmacology, metabolite generation, and drug safety.

After loading fenfluramine into CT-link, the T-link predictive module is executed and the M-link and MT-link metabolite generation and profiling modules are activated. The results of T-link are shown in Figure 2.

Every row in the spreadsheet represents either a known (K) or predicted (P) interaction. When available, the corresponding known and/or predicted affinities are also provided in –log units. If there is sufficient data in the neighbourhood of the molecule being processed, the functional effect of the compound (agonist, antagonist, inhibitor, substrate) is also provided. Also, the number and type of computational methods that led to every single prediction is shown. Most importantly, a confidence score is provided for every single predicted interactions. The confidence score depends on the number and type of methods that independently predicted the interaction, but also on the strength of the predicted affinities. In general, the larger the number of methods and the stronger the predicted affinities, the higher the confidence score is. Last but not least, the full name of the protein and a direct link to UniProt is also added to offer the user easy access to detailed information on the function of the protein and its potential role in the therapeutic action of the molecule.

Finally, one of the key features of the CT-link is the ability to generate metabolites for all the structures being processed. The methodology uses a knowledge-based approach derived from a manually curated database of 8961 metabolic transformations associated with 1791 molecules, mostly drugs extracted from bibliographic sources, representing a total of 71 chemical transformations. After an atom-by-atom superposition, the algorithm detects the chemical environment that has been transformed and encodes it using chemical descriptors. This way, every chemical transformation is described by a set of fragments, their transformations, and their corresponding chemical environments. Once a molecule is processed, if one of the chemical environments present in our database is identified in the structure of the molecule, the encoded chemical transformation is applied, with a confidence score related to the probability that the presence of this environment undergoes such transformation, based on the metabolite data contained in our database.

Once all metabolite structures that can be produced above a certain confidence threshold are generated, their target profile is predicted. Comparing the pharmacology of the original substrate molecule with those of the metabolites generated allows for detecting potential liabilities from the formation of those metabolites. An illustrative example is presented in Figure 3, in which the comparative pharmacology between fenfluramine and the six metabolites generated is provided. In the comparative pharmacology application, the green portion of the bar reflects the percentage of targets for which affinities are similar between the substrate and the metabolite; the gray portion those targets lost by the metabolies; and the red portion those targets for which the original substrate molecule did not have affinity. The latter is the focus of potential concern for safety.



Name	fenflu
Status	finished

Molecules	1
Metabolites	6

Molecule Identifier	Metabolite Identifier	Similarity
fenfluramine	fenfluramine.M6	▨🟥
fenfluramine	fenfluramine.M3	▨🟥
fenfluramine	fenfluramine.M2	🟩▨🟥
fenfluramine	fenfluramine.M5	🟩▨🟥
fenfluramine	fenfluramine.M1	🟩▨🟥
fenfluramine	fenfluramine.M4	🟩▨🟥

Figure 3: Comparative pharmacology of the predicted metabolites for fenfluramine.

As can be observed, the top metabolite (the one with the highest confidence) generated for fenfluramine is norfenfluramine (M1). By clicking on the comparative pharmacology row, a new section emerges with detailed information on the differences between the target profiles of fenfluramine and norfenfluramine. Of mention is the fact that a red bar is clearly highlighted: this alerts on the difference in affinity for the $5HT_{2B}$ receptor, for which norfenfluramine is an agonist with an affinity two orders of magnitude more potent than fenfluramine. A spreadsheet tab is also provided to inspect the exact affinity values of the two molecules for the various proteins and confirm the known or predicted functional effect.

Computational Methods

As represented in Figure 4, the system behind CT-link was developed following the classical three-tier software architecture. The data tier stores both the known and predicted data, the application tier computes the predictions and prepares data for visualizations and finally the client tier is responsible for the user interaction.

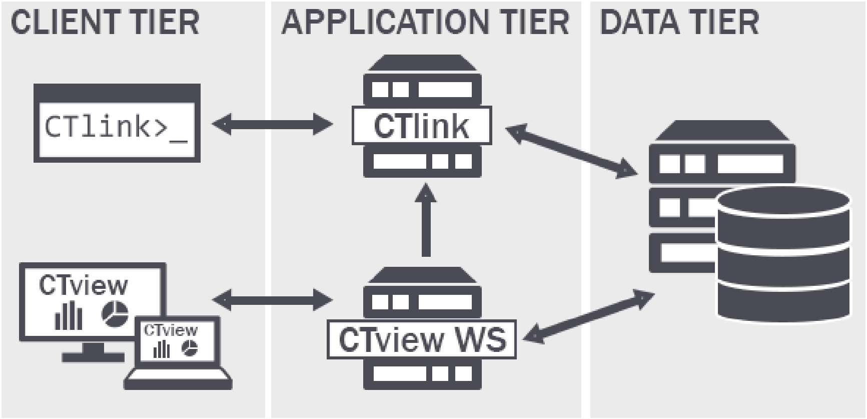


Figure 4: Three tier architecture of CT-link.

CT-link generates the prediction data starting from a series of input molecules provided by the user. Then, the application stores the data of the predictions generated in the database. As soon as the data is available, the GUI allows the user to answer the questions of interest. In technical terms, we use a relational database as an asynchronous data broker between the data predictive system and the data GUI. The data obtained is stored in a PostgreSQL database v9.6[19][20].

The client tier has been developed using Apache Flex. Apache Flex http://flex.apache.org/ is an open-source framework for building expressive web and mobile applications. Apache Flex uses MXML language for layout and AS3 language for coding. Using a set of pre-defined components, we have developped a web interface that shows data in different ways, mainly as interactive spreadsheets and charts. The application provides also tools to filter and sort the data using different criteria. Using these tools, the user can extract knowledge from the known and predicted data depending on the needs.

The web service layer is a RESTful API[21] developed using the Play framework[22] which is a Scala programming language[23][24] web framework. Scala is a general-purpose multi-paradigm programming language providing support for object-oriented and functional programming styles. It has a strong static type system with type inference. Scala runs on the Java platform (Java virtual machine) and it is compatible with all existing Java libraries.

The Play framework is based on a lightweight, stateless, web-friendly architecture and features predictable and minimal resource consumption (CPU, memory, threads) for highly-scalable applications thanks to its reactive model, based on Akka Streams[25]

References

[1] Dimasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. J Health Econ 2003;22:151–85. doi:10.1016/S0167-6296(02)00126-1.

[2] Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. Nat Publ Gr 2012;11. doi:10.1038/nrd3681.

[3] Morgan S, Grootendorst P, Lexchin J, Cunningham C, Greyson D. The cost of drug development: A systematic review. Health Policy (New York) 2010;100:4–17. doi:10.1016/j.healthpol.2010.12.002.

[4] DiMasi JA, Grabowski HG, Hansen RW. Innovation in the pharmaceutical industry: New estimates of R&amp;D costs. J Health Econ 2016;47:20–33. doi:10.1016/j.jhealeco.2016.01.012.

[5] Gaulton A, Bellis LJ, Bento AP, Chambers J, Davies M, Hersey A, et al. ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 2012;40:D1100-7. doi:10.1093/nar/gkr777.

[6] Kim S, Thiessen PA, Bolton EE, Chen J, Fu G, Gindulyte A, et al. PubChem Substance and Compound databases. Nucleic Acids Res 2016;44:D1202-13.

doi:10.1093/nar/gkv951.

[7] Wang Y, Bryant SH, Cheng T, Wang J, Gindulyte A, Shoemaker BA, et al. PubChem BioAssay: 2017 update. Nucleic Acids Res 2017;45:D955–63. doi:10.1093/nar/gkw1118.

[8] Manyika J (James), Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, et al. Big Data : The next frontier for innovation, competition, and Productivity. McKinsey Global Institute; 2011.

[9] Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. Int J Inf Manage 2015;35:137–44. doi:10.1016/j.ijinfomgt.2014.10.007.

[10] Gandomi A, Haider M. Beyond the hype: Big data concepts, methods, and analytics. Int J Inf Manage 2015;35:137–44. doi:10.1016/j.ijinfomgt.2014.10.007.

[11] Margolis R, Derr L, Dunn M, Huerta M, Larkin J, Sheehan J, et al. The National Institutes of Health's Big Data to Knowledge (BD2K) initiative: capitalizing on biomedical big data. J Am Med Inform Assoc 2014;21:957–8. doi:10.1136/amiajnl-2014-002974.

[12] Allen A, Aragon C, Becker C, Carver JC, Chis A, Combemale B, et al. Lightning talk: "I solemnly pledge" A manifesto for personal responsibility in the engineering

of academic software. CEUR Workshop Proc., vol. 1686, Nature Publishing Group; 2016, p. 160018. doi:10.1038/sdata.2016.18.

[13] Durinx C, McEntyre J, Appel R, Apweiler R, Barlow M, Blomberg N, et al. Identifying ELIXIR Core Data Resources. F1000Research 2017;5:2422. doi:10.12688/f1000research.9656.2.

[14] Garcia-Serna R, Mestres J. Anticipating drug side effects by comparative pharmacology. Expert Opin Drug Metab Toxicol 2010;6:1253–63. doi:10.1517/17425255.2010.509343.

[15] Tukey JW. Exploratory Data Analysis. Analysis 1977;2:688. doi:10.1007/978-1-4419-7976-6.

[16] Thacker C, McCreight E, Lampson B, Sproull R, Boggs D. Alto: A personal computer. Comput Struct Readings Examples 1979:549–72.

[17] Engelbart DC, English WK. A research center for augmenting human intellect. Proc. December 9-11, 1968, fall Jt. Comput. Conf. part I - AFIPS '68 (Fall, part I), New York, New York, USA: ACM Press; 1968, p. 395. doi:10.1145/1476589.1476645.

[18] Heer J, Shneiderman B. Interactive dynamics for visual analysis. Commun ACM 2012;55:45.

doi:10.1145/2133806.2133821.

[19] PostgreSQL. PostgreSQL 2016.

[20] Stonebraker M, Rowe LA. The design of POSTGRES. Proc. 1986 ACM SIGMOD Int. Conf. Manag. data - SIGMOD '86, 1986, p. 340–55. doi:10.1145/16894.16888.

[21] Fielding, Thomas R. Architectural styles and the design of network-based software architectures. University of California, Irvine; 2000.

[22] Hilton P, Bakker E, Canedo F. Play for Scala : covers Play 2. Manning; 2014.

[23] Odersky M, Altherr P, Cremet V, Emir B, Maneth S, Micheloud S, et al. An Overview of the Scala Programming Language. System 2004:1–130. doi:10.1145/1706356.1706358.

[24] Odersky M, Spoon L, Venners B. Programming in Scala: a comprehensive step-by-step guide. artima. 2010. doi:10.1016/j.tcs.2008.09.019.

[25] Roestenburg R, Bakker R, Williams R. Akka in action. Manning; 2015.

3.6.1 TECHNICAL RESULTS

In the following sections, we are going to explain the technical details of the CT-link platform that are not included in the paper due to its characteristics.

3.6.1.1 FUNCTIONALITY

The CT-link platform provides the following functionalities:

- Create jobs: a job is defined by a set of molecules (with a maximum of 100). To create the job, the user can load multiple sdf or smi files, draw molecules, modify existing molecules and delete molecules from the set. That means that the platform provides whole capabilities to edit molecules sets.
- Execute jobs: given a set of molecules the user can be able to request four different types of executions: T-link (target profiling), S-link (safety profiling), M-link (metabolite structure predictions) and MT-link (target profiling of metabolites) or any combination of them.
- Explore the results using interactive charts and the classical spreadsheet visualization.
- Detailed information about the substantiation of the interactions.
- Apply filters and sort the results.
- Download generated data as CSV files or pdf reports with the substantiation information.

3.6.1.2 DESIGN

The application is designed following the classical three layer architecture: presentation layer, application layer, and data layer. In the following sections, a brief description of each layer is provided.

3.6.1.2.1 Presentation layer

The presentation layer has been implemented using Flex. Applications developed using Flex assures Rich User Experience through intuitive interaction with the application and presenting information in a visually rich interface.

In the following diagram, you can see the main modules of the presentation layer.

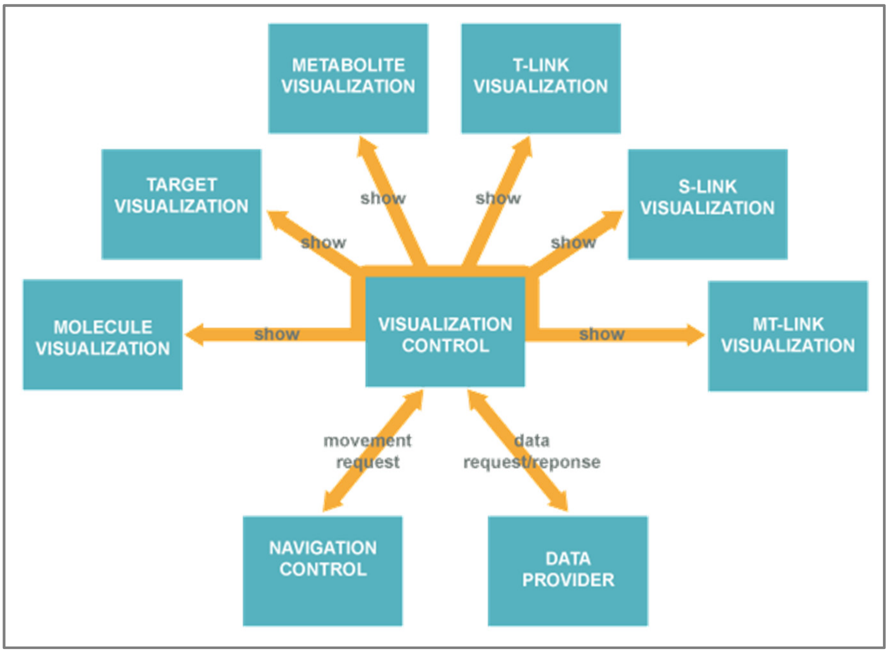


Figure 24: CTlink presentation layer main modules.

- **Navigation control**: This module controls the movements that the user can perform. This navigation control has been implemented as a network Chart. This network has four nodes, namely, molecules, targets, safety and metabolites, and four connections between these nodes, that is molecule-target, molecule-safety, molecule-metabolite, metabolite-target. The user can visit the nodes and the connections and this module controls which position is visiting the user at each moment. According to the user interactions, the module sends a message to the Visualization Control with the request for the data that has to be visualized according to the interaction with the network chart.

- **Visualization control**: This module controls which data has to be visualized according to the requests made by the Navigation Control. The module controls which entity is visualized and also which type of visualizations have to be shown. As you can see in the chart, each entity (Molecule, Target, Safety, Metabolite) has its own visualization module. And each of these modules have different ways to visualize its content. The module also has a connection with the Data Provider module to request the data that needs to show according to the user navigation and the filters applied.

- **Data Provider**: This module calls for all the data for a job and stores it. When it receives a request from the

“Visualization Control”, it returns the requested data in the requested format and with the requested filters applied.

The dynamic relation of these modules generates an interactive experience to the user.

3.6.1.2.2 Application layer

The application layer has two main parts. The prediction system and the web services that provide the data to the presentation layer. The prediction system is out of the scope of this thesis and thus, focus will be given to the web services.

Web services are developed using Scala [40] and the Play web framework [41]. The main advantage of the Play framework is its usability from the developer perspective, which makes building web applications rather straightforward.

As mentioned above, this layer has two components: the prediction system and the web services. One important point of this architecture is the way that these two components interact. To create a communication between both modules we use the database as shown in the following figure.

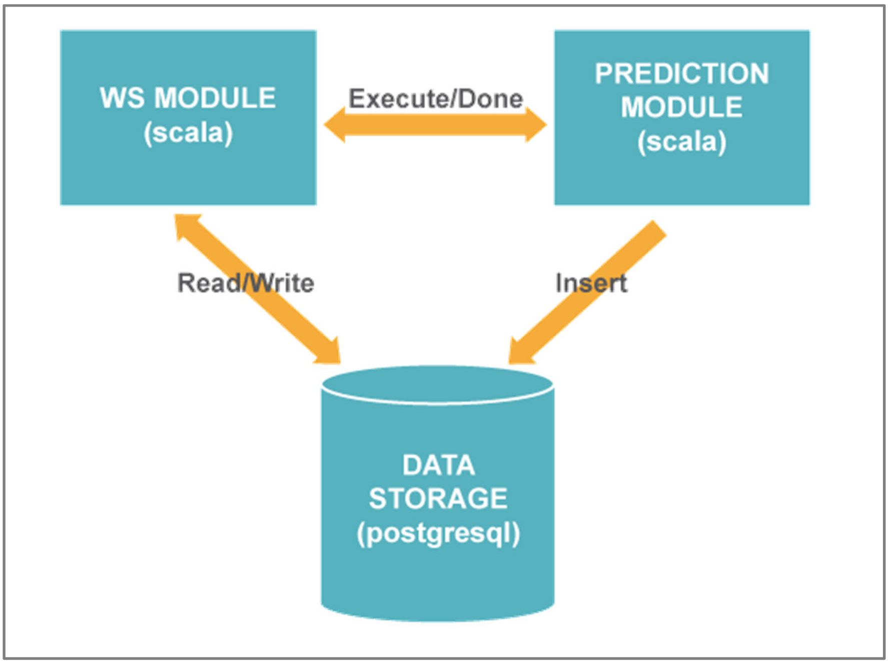


Figure 25: CT-link application layer modules communication.

The Scala module receives the request for an execution from the presentation layer. It starts an asynchronous execution of the prediction module, after which the prediction module inserts the data generated into the database and also inserts information about the progress. When the execution is finished, the Scala module recovers the control and performs some final steps. During the execution process, the Scala module can receive requests from the presentation layer about the status of the job. The Scala module can answer to these requests because it shares the database. The prediction module updates the state and the progress of the job and the Scala module can get this information to answer the presentation layer request.

We illustrate this behaviour in the following UML sequence diagram.

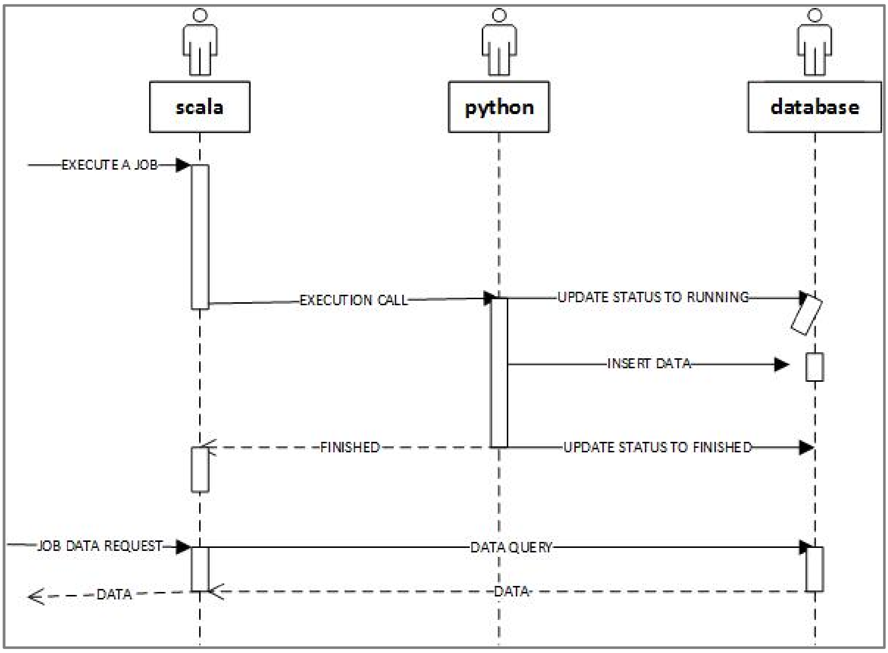


Figure 26: CT-link application layer modules communication UML.

3.6.1.2.3 Data model

The following chart shows the data model of the applications. This data model stores the predicted data generated by the system, as well as information about the users of the application and the executed jobs.

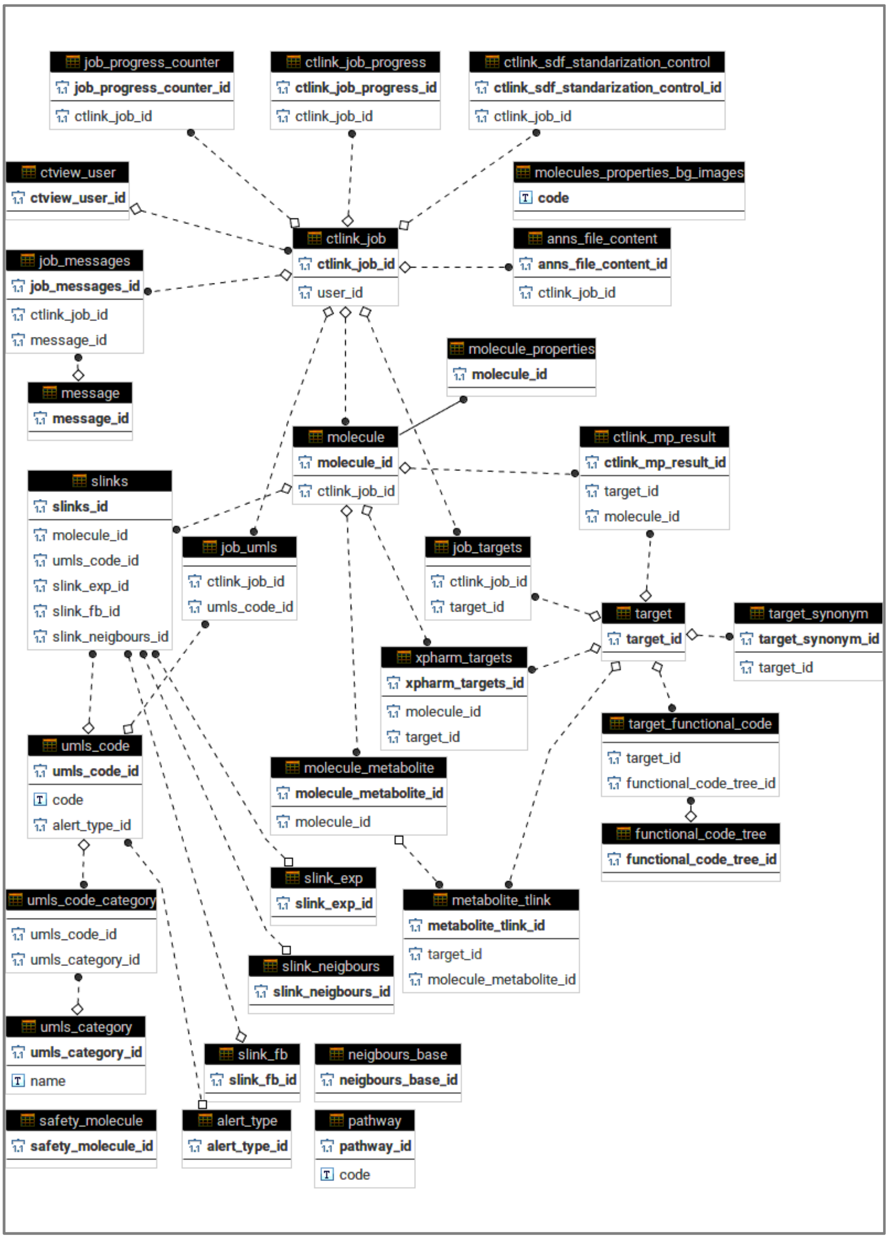


Figure 27: CT-link data model.

The data model main tables are:

CT-link_job	**This table stores the jobs executed by the different users of the application. It stores information about the execution time and the status of the execution.**
molecule	This table contains all the molecules uploaded to the application. For each molecule, the table stores the job where it belongs.
CT-link_mp_result	For each molecule, it contains the results generated by the T-link execution. This table connects molecules with targets with a certain activity value and confidence score.
slinks	For each molecule, this table contains the results of the S-link execution. This table connects molecule with uml safety codes with a certain confidence score.

molecule-metabolite	This table contains the results of the M-link execution. It connects molecules with metabolites with a certain score.
metabolite-tlink	This table contains the results of the MT-link. For each metabolite, it stores the T-link execution results.

3.6.1.3 GUI

In order to access CT-link, login authentication is required, as you can see in Figure 28.

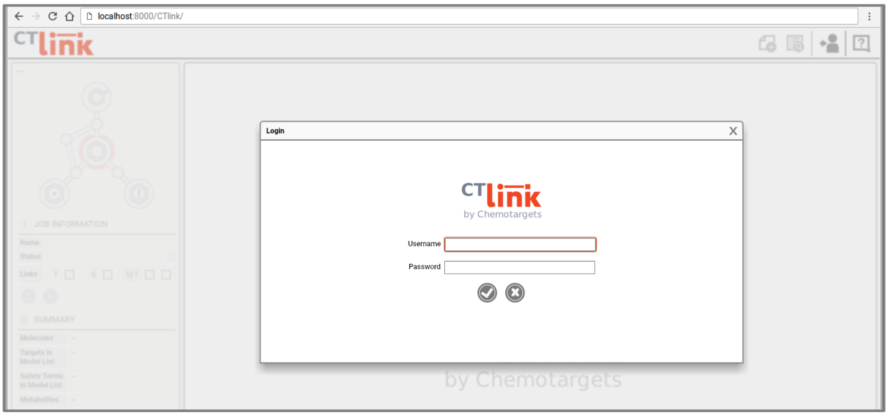


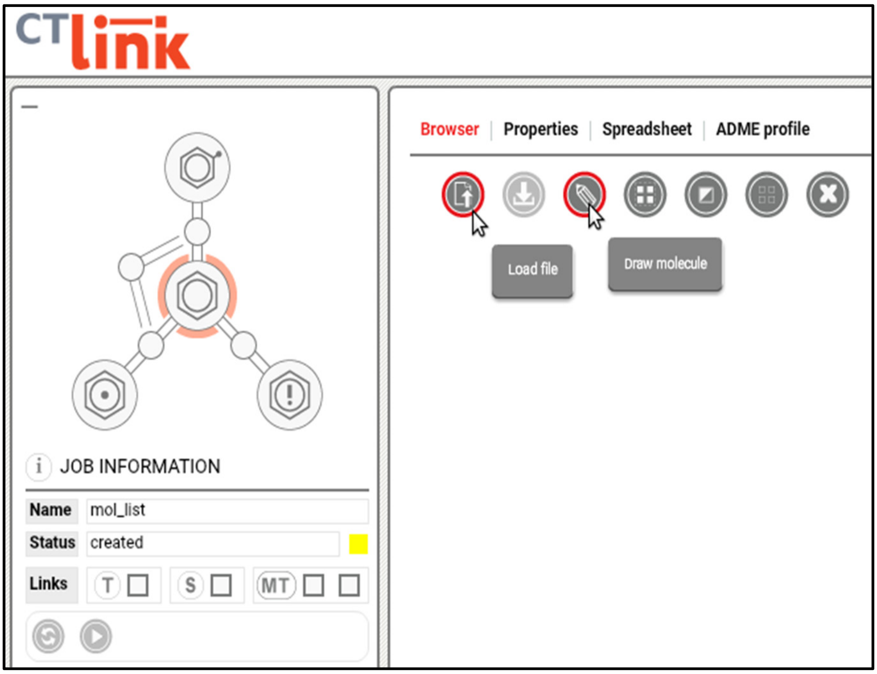
Figure 28: CT-link login screen.

After the login, the user gets directed to the main page of the application. The screen is divided into two main areas. On the left, we have the navigation panel and execution area and, on the right, the results area.

At this stage, the user can perform two tasks: create a new job or explore previous jobs.

In order to start a new CT-link job, a name for the job is required.

After entering the job name and accept, the job is created and the user can load a set of molecules that will be the input of the execution process. The user has two options to create this set of molecules: load multiple sdf or smi files or draw molecules (Figure 29).



JOB INFORMATION

Name	mol_list
Status	created 🟨
Links	T □ S □ MT □ □

Figure 29: CT-link job creation screen.

Once loaded, molecules can be easily added, modified or deleted.

After loading the molecules, the user can execute a job. To run a job, the user has to select the execution that will perform: T-link, S-link, M-link or MT-link. The user can select multiple options at the same time.

After selecting the execution type, the calculation can be started using the "Run job" button.

The user can control the status of the job during all the process. A progress bar appears to show an estimation of the remaining execution time.

When the job is finished, the user can explore the results using the navigation panel (Figure 31).

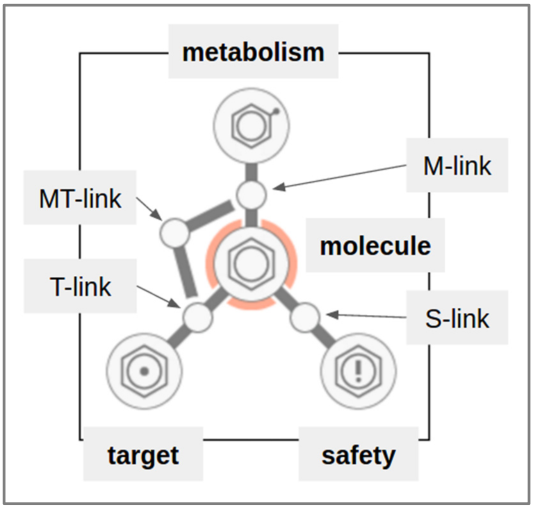


Figure 31: CTlink navigation control.

The current CT-link application has four different nodes, one central -Molecule- connected to three peripheral -Target, Safety and Metabolism- connected by three different *links*; T-link (Molecule-Target), S-link (Molecule-Safety) and M-link (Molecule-Metabolism). In addition, one extra link called M[T]-link, which connects the T-link and the M-link, has been specifically designed to compare the predicted primary target profiles of a given parent molecule (substrate) with their corresponding predicted metabolites.

3.6.1.3.1 Molecules node

The molecules entity has several tabs to explore the data in different ways. In the "Browser" tab the user can see the molecules as cards with an image of the molecule, the name and the inchiKey (Figure 32).

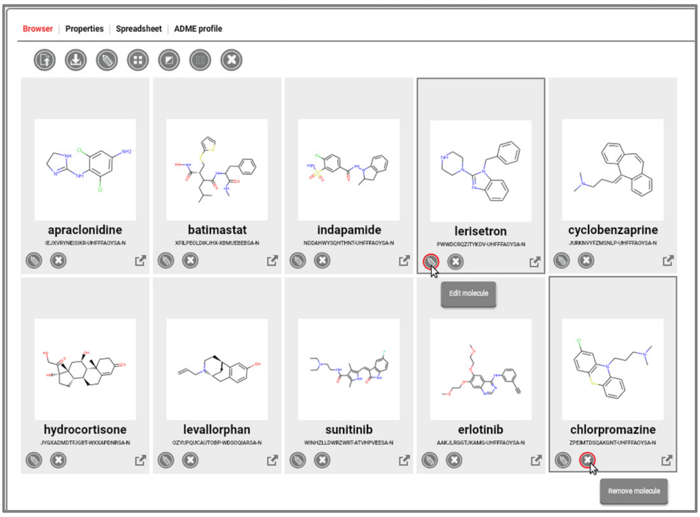
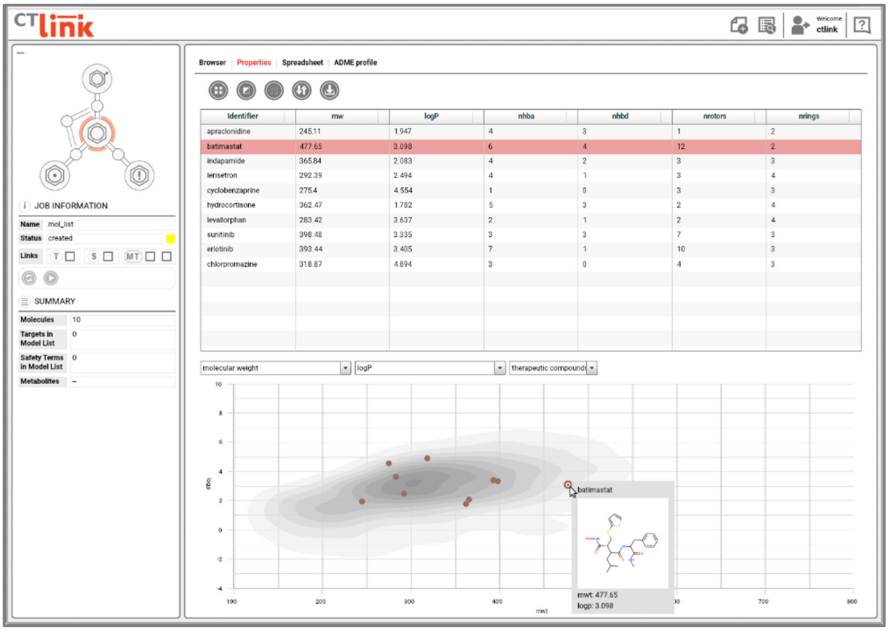

Browser
Properties
Spreadsheet
ADME profile
apraclonidine
IEJXVRYNEISIKR-UHFFFAOYSA-N
batimastat
XFILPEOLDIKJHX-XBMUEBEBSA-N
indapamide
NDDAHWYSQHTHNT-UHFFFAOYSA-N
lerisetron
PWWDCRQZITYKDV-UHFFFAOYSA-N
cyclobenzaprine
JURKNVYFZMSNLP-UHFFFAOYSA-N
Edit molecule
hydrocortisone
JYGXADMDTFJGBT-WXXAPDNRSA-N
levallorphan
OZYUPQUCAUTOBP-WDSOQIARSA-N
sunitinib
WINHZLLDWRZWRT-ATVHPVEESA-N
erlotinib
AAKJLRGGTJKAMG-UHFFFAOYSA-N
chlorpromazine
ZPEIMTDSQAKGNT-UHFFFAOYSA-N
Remove molecule

Figure 32: CTlink molecule cards.

In the “Properties” tab, several molecular properties are represented as a bidimensional chart with different reference backgrounds: therapeutic compounds (drugs), bioactive compounds, natural products and synthetic compounds (Figure 33).



Name: mol_list
Status: created
Links: T □ S □ MT □ □

Molecules: 10
Targets in Model List: 0
Safety Terms in Model List: 0
Metabolites: –

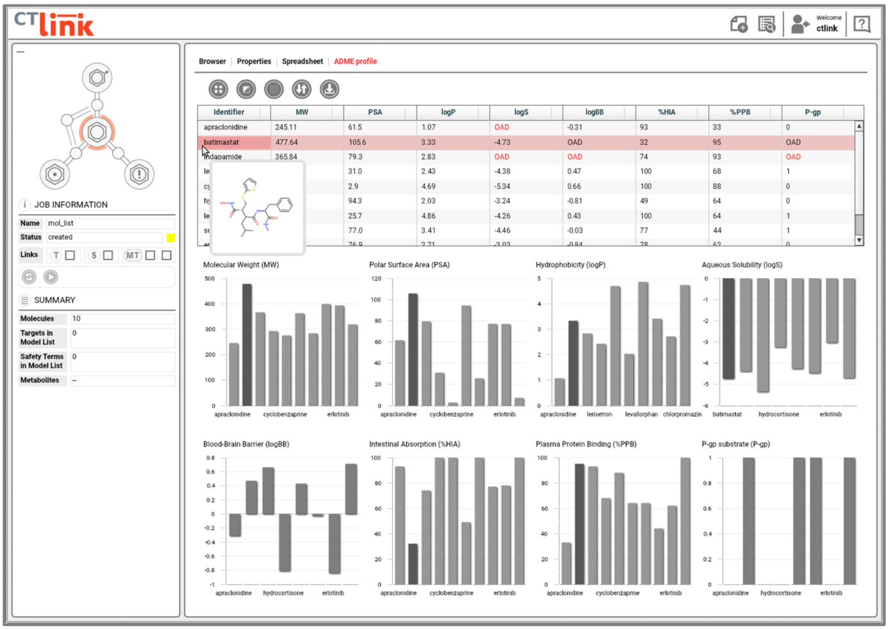
Identifier	mw	logP	nhba	nhbd	nrotors	nrings
apraclonidine	245.11	1.947	4	3	1	2
batimastat	477.65	3.098	6	4	12	2
indapamide	365.84	2.083	4	2	3	3
lerisetron	292.39	2.494	4	1	3	4
cyclobenzaprine	275.4	4.554	1	0	3	3
hydrocortisone	362.47	1.782	5	3	2	4
levallorphan	283.42	3.637	2	1	2	4
sunitinib	398.48	3.335	3	3	7	3
erlotinib	393.44	3.405	7	1	10	3
chlorpromazine	318.87	4.894	3	0	4	3

molecular weight | logP | therapeutic compounds

mwt: 477.65
logp: 3.098

Figure 33: CT-link properties tab.

In the “Spreadsheet” tab, the molecules are shown as rows in a spreadsheet. And in the “ADME profile”, a group of different ADME-related properties are predicted using fragment contribution-based QSPR models, which have been exclusively developed at Chemotargets using experimental data collected from different sources (Figure 34).



JOB INFORMATION

Name	mol_list
Status	created
Links	T □ S □ MT □ □

SUMMARY

Molecules	10
Targets in Model List	0
Safety Terms in Model List	0
Metabolites	--

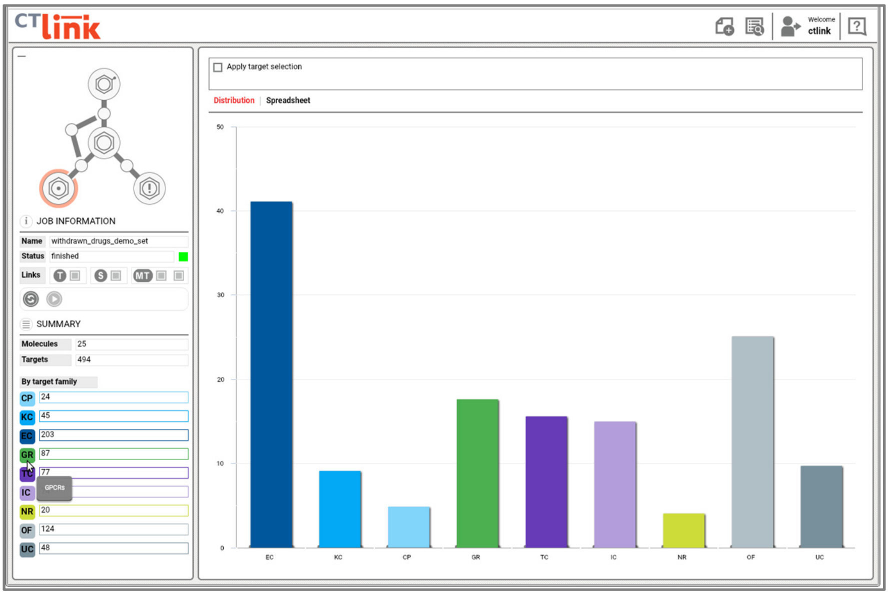
Identifier	MW	PSA	logP	logS	logBB	%HIA	%PPB	P-gp
apraclonidine	245.11	61.5	1.07	OAD	-0.31	93	33	0
batimastat	477.64	105.6	3.33	-4.73	OAD	32	95	OAD
indapamide	365.84	79.3	2.83	OAD	OAD	74	93	OAD
le		31.0	2.43	-4.38	0.47	100	68	1
cy		2.9	4.69	-5.34	0.66	100	88	0
hy		94.3	2.03	-3.24	-0.81	49	64	0
le		25.7	4.86	-4.26	0.43	100	64	1
su		77.0	3.41	-4.46	-0.03	77	44	1
er		76.9	2.71	-3.03	-0.84	78	62	0

Figure 34: CT-link molecules ADME profile tab.

3.6.1.3.2 Target Node

When the user clicks on the target node of the navigation panel, a screen with two tabs appears: “Distribution” and “Spreadsheet”.

In the Distribution tab, the user can see the targets of the jobs in a bar chart where each bar represents a target family and the bar’s height represents the number of targets of each family (Figure 35).



JOB INFORMATION

Name	withdrawn_drugs_demo_set
Status	finished

SUMMARY

Molecules	25
Targets	494

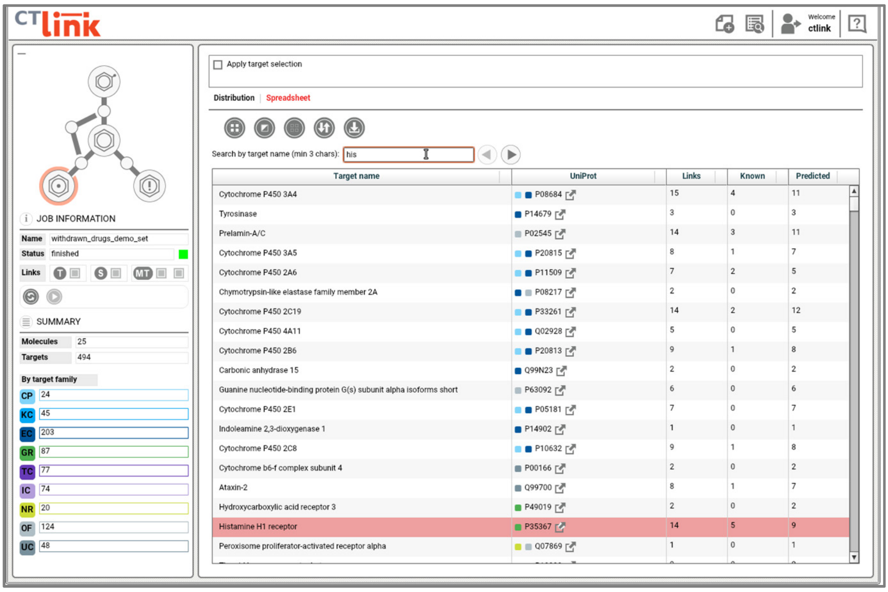
By target family

CP	24
KC	45
EC	203
GR	87
TC	77
IC	[illegible]
NR	20
OF	124
UC	48

☐ Apply target selection

Figure 35: CT-link target node "Distribution" tab.

In the Spreadsheet tab, the user can see a summary of the number of *links* identified between the input molecules and the different targets and target families. The user can easily find the target of interest using the search field (Figure 36).



JOB INFORMATION

Name	withdrawn_drugs_demo_set
Status	finished

SUMMARY

Molecules	25
Targets	494

By target family

CP	24
KC	45
EC	203
GR	87
TC	77
IC	74
NR	20
OF	124
UC	48

□ Apply target selection

Distribution | Spreadsheet

Search by target name (min 3 chars): his

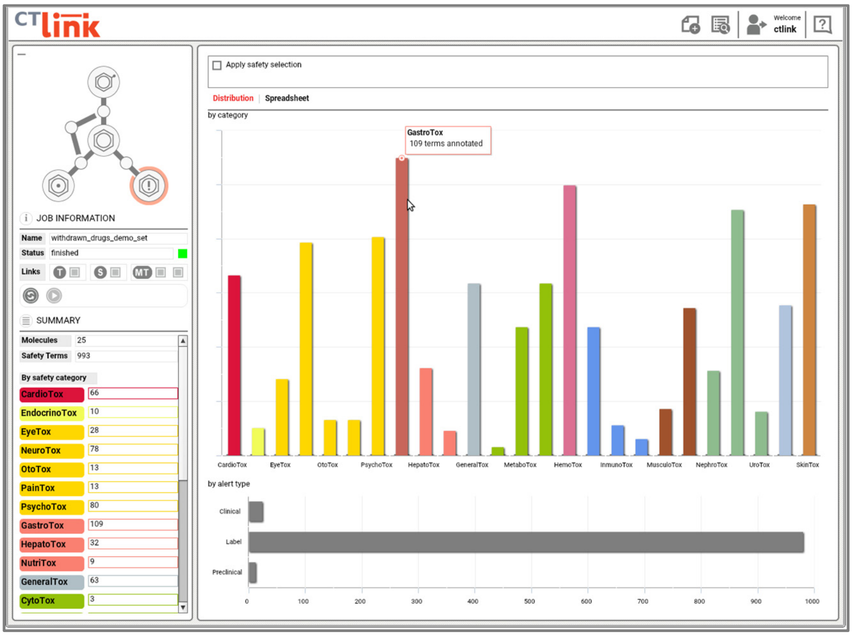
Target name	UniProt	Links	Known	Predicted
Cytochrome P450 3A4	P08684	15	4	11
Tyrosinase	P14679	3	0	3
Prelamin-A/C	P02545	14	3	11
Cytochrome P450 3A5	P20815	8	1	7
Cytochrome P450 2A6	P11509	7	2	5
Chymotrypsin-like elastase family member 2A	P08217	2	0	2
Cytochrome P450 2C19	P33261	14	2	12
Cytochrome P450 4A11	Q02928	5	0	5
Cytochrome P450 2B6	P20813	9	1	8
Carbonic anhydrase 15	Q99N23	2	0	2
Guanine nucleotide-binding protein G(s) subunit alpha isoforms short	P63092	6	0	6
Cytochrome P450 2E1	P05181	7	0	7
Indoleamine 2,3-dioxygenase 1	P14902	1	0	1
Cytochrome P450 2C8	P10632	9	1	8
Cytochrome b6-f complex subunit 4	P00166	2	0	2
Ataxin-2	Q99700	8	1	7
Hydroxycarboxylic acid receptor 3	P49019	2	0	2
Histamine H1 receptor	P35367	14	5	9
Peroxisome proliferator-activated receptor alpha	Q07869	1	0	1

Figure 36: CT-link target node "Spreadsheet" tab.

3.6.1.3.3 Safety node

The results area for the safety node selection also presents two tabs: “Distribution” and “Spreadsheet”.

Under the Distribution tab, the user can see a coloured vertical bar chart where each bar represents the number of endpoints of each category that have at least one annotation. Also, a grey horizontal bars is shown. In this chart, each bar represents the number of endpoints of each alert type that has at least one annotation. All these charts are interactive, and the user can click on any bar and apply a selection including all related terms (Figure 37).



Name	withdrawn_drugs_demo_set
Status	finished

Molecules	25
Safety Terms	993

By safety category

CardioTox	66
EndocrinoTox	10
EyeTox	28
NeuroTox	78
OtoTox	13
PainTox	13
PsychoTox	80
GastroTox	109
HepatoTox	32
NutriTox	9
GeneralTox	63
CytoTox	3

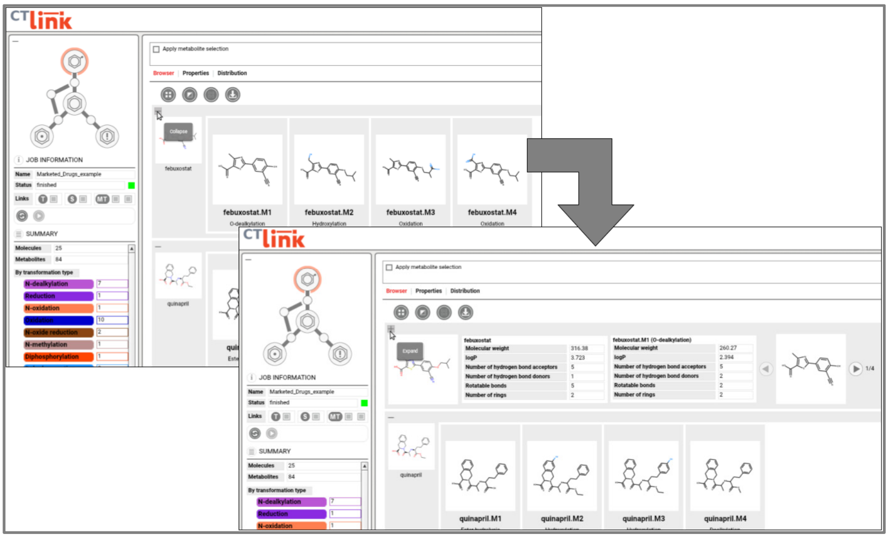
Figure 37: CT-link safety node "Distribution" tab.

The “Spreadsheet” tab shows a summary of the number of *links* identified between the input molecules and the different safety terms and categories.

3.6.1.3.4 Metabolism node

The metabolism results area has three tabs: “Browser”, “Properties” and “Distribution”.

The “Browser” tab shows, for each parent molecule, the structure and name of the generated metabolites along with the name of the corresponding metabolic transformation. These panels can be easily collapsed and expanded as you can see in Figure 38.



Name	Marketed_Drugs_example
Status	finished

Molecules	25
Metabolites	84

By transformation type

N-dealkylation	7
Reduction	1
N-oxidation	1
Oxidation	10
N-oxide reduction	2
N-methylation	1
Diphosphorylation	1

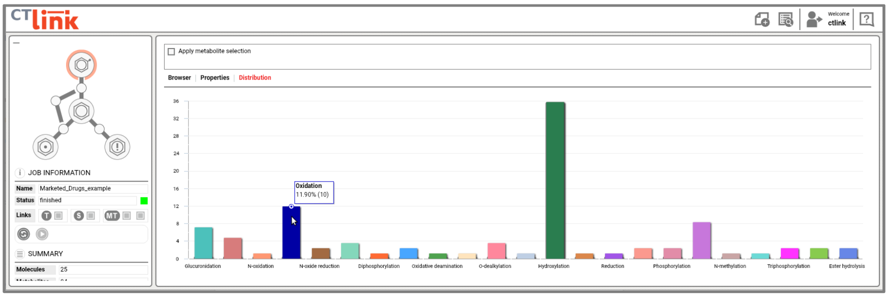
febuxostat	
Molecular weight	316.38
logP	3.723
Number of hydrogen bond acceptors	5
Number of hydrogen bond donors	1
Rotatable bonds	5
Number of rings	2

febuxostat.M1 (O-dealkylation)	
Molecular weight	260.27
logP	2.394
Number of hydrogen bond acceptors	5
Number of hydrogen bond donors	2
Rotatable bonds	2
Number of rings	2

Figure 38: CT-link metabolite node "Browser" tab.

The “Properties” tab, similarly to the *“Molecule”* node, represents several molecular properties as a bidimensional chart with different reference backgrounds: therapeutic compounds (drugs), bioactive compounds, natural products and synthetic compounds.

The “Distribution” tab, display vertical coloured bars that represent the number of metabolites generated as a function of the different metabolic transformations (Figure 39).



□ Apply metabolite selection

JOB INFORMATION

Name	Marketed_Drugs_example
Status	finished

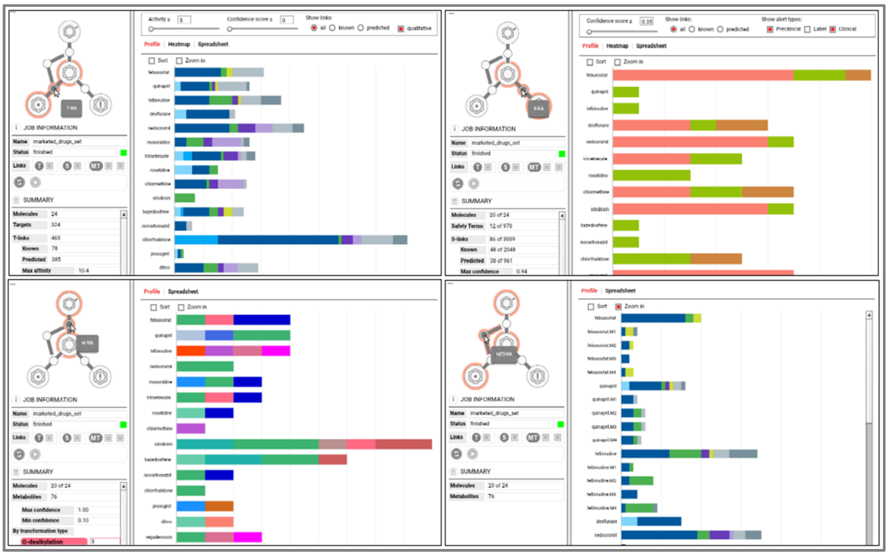
SUMMARY

Molecules	25

Figure 39: CT-link metabolite node "Distribution" tab.

3.6.1.3.5 Links

The different links share some visualization modes. All links have a “Profile” tab where the different profiles of screened molecules are represented by a stacked bars chart (Figure 40). For each stacked bar representing a molecule, the length represents the number of *links* to every target family. The bars can be sorted and/or zoom view applied by clicking on the corresponding checkboxes.



Activity ≥ 5 Confidence score ≥ 0 Show links: ● all ○ known ○ predicted ☒ qualitative

☐ Sort ☐ Zoom in

Name marketed_drugs_set
Status finished

Molecules 24
Targets 324
T-links 463
Known 78
Predicted 385
Max affinity 10.4

Confidence score ≥ 0.35 Show links: ● all ○ known ○ predicted Show alert types: ☒ Preclinical ☐ Label ☒ Clinical

☐ Sort ☐ Zoom in

Name marketed_drugs_set
Status finished

Molecules 20 of 24
Safety Terms 12 of 970
S-links 86 of 3009
Known 48 of 2048
Predicted 38 of 961
Max confidence 0.94

☐ Sort ☐ Zoom in

Name marketed_drugs_set
Status finished

Molecules 20 of 24
Metabolites 76
Max confidence 1.00
Min confidence 0.10
By transformation type
O-dealkylation 3

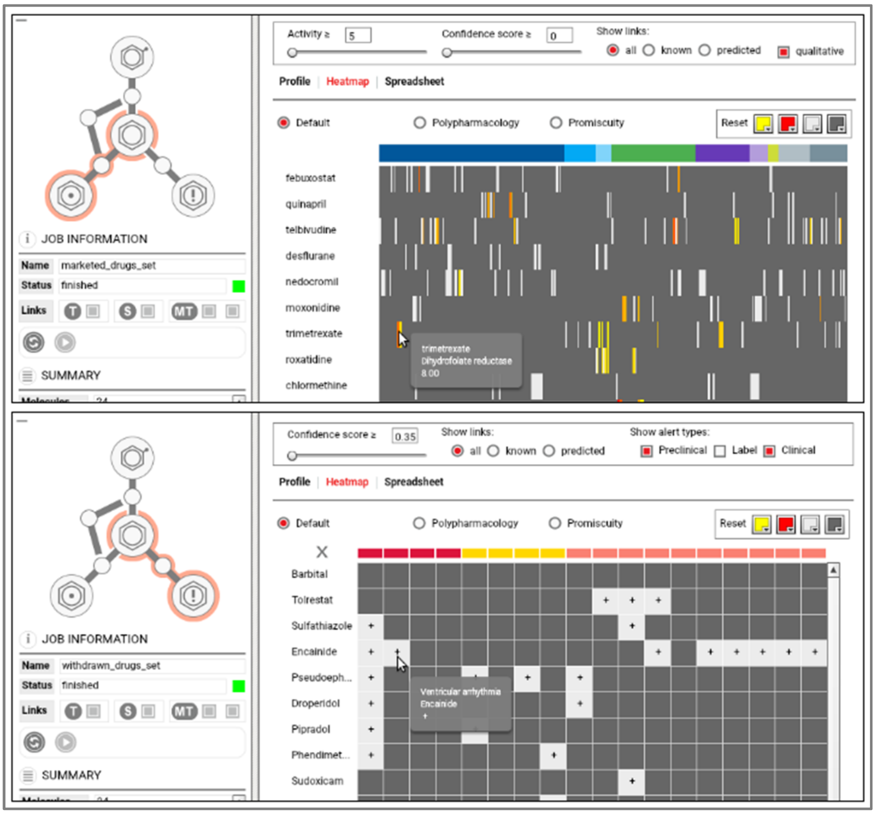
☐ Sort ☒ Zoom in

Name marketed_drugs_set
Status finished

Molecules 20 of 24
Metabolites 76

Figure 40: CT-link edges "Profile" tab.

For the T-link and S-link results a “Heatmap” tab is also available. The heatmaps can be easily explored by double-clicking in the region of interest to zoom-in (Figure 41).



JOB INFORMATION

Name marketed_drugs_set

Status finished

Links

SUMMARY

Activity ≥ 5

Confidence score ≥ 0

Show links: ● all ○ known ○ predicted ☒ qualitative

Profile | Heatmap | Spreadsheet

● Default ○ Polypharmacology ○ Promiscuity Reset

febuxostat
quinapril
telbivudine
desflurane
nedocromil
moxonidine
trimetrexate
roxatidine
chlormethine

trimetrexate
Dihydrofolate reductase
8.00

JOB INFORMATION

Name withdrawn_drugs_set

Status finished

Links

SUMMARY

Confidence score ≥ 0.35

Show links: ● all ○ known ○ predicted

Show alert types: ☒ Preclinical □ Label ☒ Clinical

Profile | Heatmap | Spreadsheet

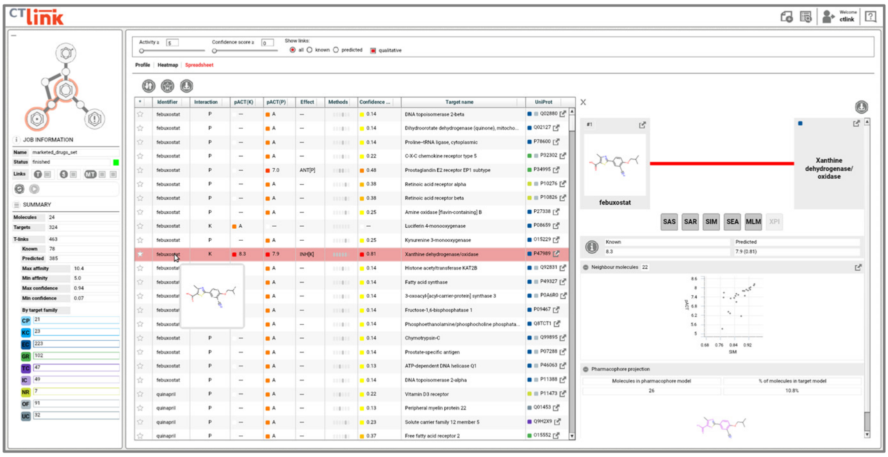
● Default ○ Polypharmacology ○ Promiscuity Reset

Barbital
Tolrestat
Sulfathiazole
Encainide
Pseudoeph...
Droperidol
Pipradol
Phendimet...
Sudoxicam

Ventricular arrhythmia
Encainide
+

Figure 41: CT-link heatmaps.

And for all the links a “Spreadsheet” tab is also provided. The spreadsheets contain the ultimate data results generated by CT-link. Additionally, by clicking on a given row, prediction substantiation is displayed at the right side of the “Results panel”. Spreadsheet contents are, obviously, *links*-specific. For T-link includes substantiation on neighbour target ligands, pharmacophore projection based on SAS, and cross-pharmacologically related targets (Figure 42).



Name	marketed_drugs_set
Status	finished
Links	T S MT

Molecules	24
Targets	324
T-links	463
Known	78
Predicted	385
Max affinity	10.4
Min affinity	5.0
Max confidence	0.94
Min confidence	0.07
By target family	
CP	21
KC	23
EC	223
GR	102
TC	47
IC	49
NR	7
OF	91
UC	32

Activity ≥ 5 Confidence score ≥ 0

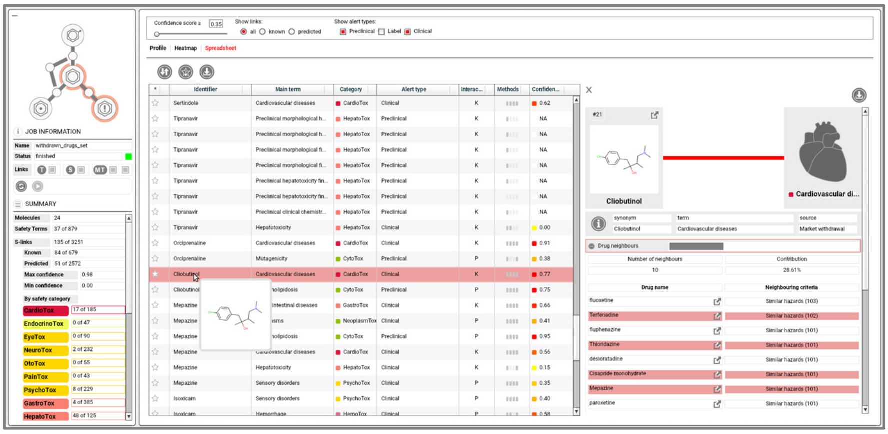
*	Identifier	Interaction	pACT(K)	pACT(P)	Effect	Methods	Confidence ...	Target name	UniProt
☆	febuxostat	P	—	A	—		0.14	DNA topoisomerase 2-beta	Q02880
☆	febuxostat	P	—	A	—		0.14	Dihydroorotate dehydrogenase (quinone), mitocho...	Q02127
☆	febuxostat	P	—	A	—		0.14	Proline--tRNA ligase, cytoplasmic	P78600
☆	febuxostat	P	—	A	—		0.22	C-X-C chemokine receptor type 5	P32302
☆	febuxostat	P	—	7.0	ANT[P]		0.48	Prostaglandin E2 receptor EP1 subtype	P34995
☆	febuxostat	P	—	A	—		0.38	Retinoic acid receptor alpha	P10276
☆	febuxostat	P	—	A	—		0.38	Retinoic acid receptor beta	P10826
☆	febuxostat	P	—	A	—		0.25	Amine oxidase [flavin-containing] B	P27338
☆	febuxostat	K	A	—	—		—	Luciferin 4-monooxygenase	P08659
☆	febuxostat	P	—	A	—		0.25	Kynurenine 3-monooxygenase	O15229
★	febuxostat	K	8.3	7.9	INH[K]		0.81	Xanthine dehydrogenase/oxidase	P47989
☆	febuxostat			A	—		0.14	Histone acetyltransferase KAT2B	Q92831
☆	febuxostat			A	—		0.14	Fatty acid synthase	P49327
☆	febuxostat			A	—		0.14	3-oxoacyl-[acyl-carrier-protein] synthase 3	P0A6R0
☆	febuxostat			A	—		0.14	Fructose-1,6-bisphosphatase 1	P09467
☆	febuxostat			A	—		0.14	Phosphoethanolamine/phosphocholine phosphata...	Q8TCT1
☆	febuxostat	P	—	A	—		0.14	Chymotrypsin-C	Q99895
☆	febuxostat	P	—	A	—		0.14	Prostate-specific antigen	P07288
☆	febuxostat	P	—	A	—		0.13	ATP-dependent DNA helicase Q1	P46063
☆	febuxostat	P	—	A	—		0.14	DNA topoisomerase 2-alpha	P11388
☆	quinapril	P	—	A	—		0.22	Vitamin D3 receptor	P11473
☆	quinapril	P	—	A	—		0.13	Peripheral myelin protein 22	Q01453
☆	quinapril	P	—	A	—		0.23	Solute carrier family 12 member 5	Q9H2X9
☆	quinapril	P	—	A	—		0.37	Free fatty acid receptor 2	O15552

Known	Predicted
8.3	7.9 (0.81)

Molecules in pharmacophore model	% of molecules in target model
26	10.8%

Figure 42: CT-link Tlink spreadsheet and substantiation.

Molecule-Safety connections include substantiations providing information on drug neighbours, molecular fragments and related targets and pathways. A red background indicates that the neighbour drug was withdrawn from the market for producing that specific safety event (Figure 43).



Name	withdrawn_drugs_set
Status	finished
Links	T S MT

Molecules	24
Safety Terms	37 of 879
S-links	135 of 3251
Known	84 of 679
Predicted	51 of 2572
Max confidence	0.98
Min confidence	0.00
By safety category	
CardioTox	17 of 185
EndocrinoTox	0 of 47
EyeTox	0 of 90
NeuroTox	2 of 232
OtoTox	0 of 55
PainTox	0 of 43
PsychoTox	8 of 229
GastroTox	4 of 385
HepatoTox	48 of 125

Confidence score ≥ 0.35

*	Identifier	Main term	Category	Alert type	Interac...	Methods	Confiden...
☆	Sertindole	Cardiovascular diseases	CardioTox	Clinical	K		0.62
☆	Tipranavir	Preclinical morphological h...	HepatoTox	Preclinical	K		NA
☆	Tipranavir	Preclinical morphological h...	HepatoTox	Preclinical	K		NA
☆	Tipranavir	Preclinical morphological fi...	HepatoTox	Preclinical	K		NA
☆	Tipranavir	Preclinical morphological fi...	HepatoTox	Preclinical	K		NA
☆	Tipranavir	Preclinical hepatotoxicity fin...	HepatoTox	Preclinical	K		NA
☆	Tipranavir	Preclinical hepatotoxicity fin...	HepatoTox	Preclinical	K		NA
☆	Tipranavir	Preclinical clinical chemistr...	HepatoTox	Preclinical	K		NA
☆	Tipranavir	Hepatotoxicity	HepatoTox	Clinical	K		0.00
☆	Orciprenaline	Cardiovascular diseases	CardioTox	Clinical	K		0.91
☆	Orciprenaline	Mutagenicity	CytoTox	Preclinical	P		0.38
★	Cliobutinol	Cardiovascular diseases	CardioTox	Clinical	K		0.77
☆	Cliobutinol	holipidosis	CytoTox	Preclinical	P		0.75
☆	Mepazine	intestinal diseases	GastroTox	Clinical	K		0.66
☆	Mepazine	sms	NeoplasmTox	Clinical	P		0.41
☆	Mepazine	holipidosis	CytoTox	Preclinical	P		0.95
☆	Mepazine	Cardiovascular diseases	CardioTox	Clinical	K		0.56
☆	Mepazine	Hepatotoxicity	HepatoTox	Clinical	K		0.15
☆	Mepazine	Sensory disorders	PsychoTox	Clinical	P		0.35
☆	Isoxicam	Sensory disorders	PsychoTox	Clinical	P		0.40
☆	Isoxicam	Hemorrhage	HemoTox	Clinical	P		0.58

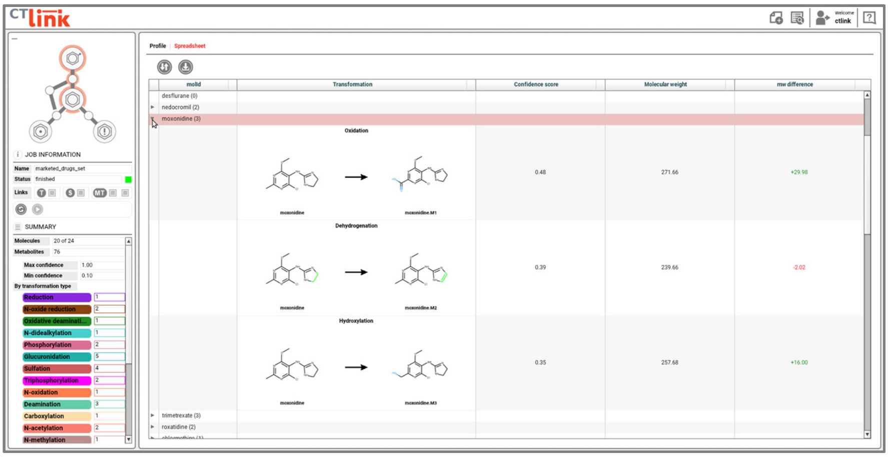
synonym	term	source
Cliobutinol	Cardiovascular diseases	Market withdrawal

Number of neighbours	Contribution
10	28.61%

Drug name	Neighbouring criteria
fluoxetine	Similar hazards (103)
Terfenadine	Similar hazards (102)
fluphenazine	Similar hazards (101)
Thioridazine	Similar hazards (101)
desloratadine	Similar hazards (101)
Cisapride monohydrate	Similar hazards (101)
Mepazine	Similar hazards (101)
paroxetine	Similar hazards (101)

Figure 43: CT-link Slink spreadsheet and substantiation.

In the M-link spreadsheet tab, by clicking on a given molecule row the user can unfold the full list of predicted metabolic transformations, with the corresponding molecular depictions, sorted by confidence score. Transformations are mapped on the molecular depictions in a color-coded fashion: red for atom deletions (in substrate), blue for atom additions (in metabolite) and green for modifications in bond order (both substrate and metabolite) (Figure 44).



JOB INFORMATION

Name	marketed_drugs_set
Status	finished 🟩

SUMMARY

Molecules	20 of 24
Metabolites	76
Max confidence	1.00
Min confidence	0.10

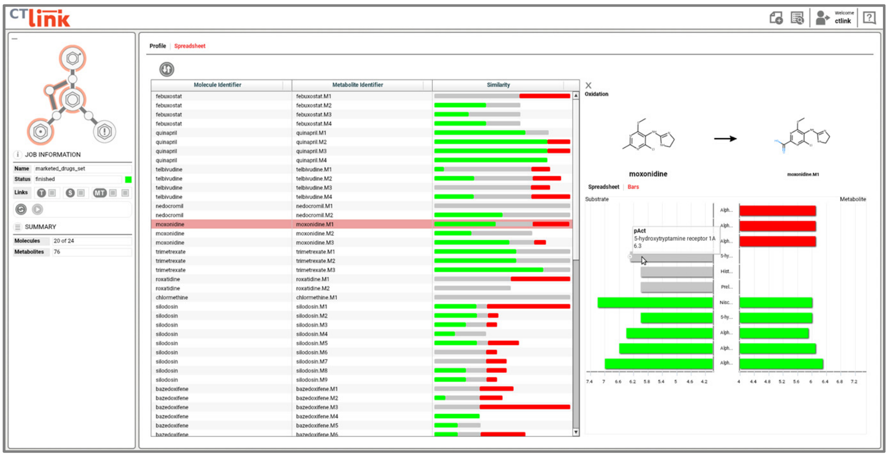
By transformation type

Reduction	1
N-oxide reduction	2
Oxidative deaminati...	1
N-didealkylation	1
Phosphorylation	2
Glucuronidation	5
Sulfation	4
Triphosphorylation	2
N-oxidation	1
Deamination	3
Carboxylation	1
N-acetylation	2
N-methylation	1

molId	Transformation	Confidence score	Molecular weight	mw difference
desflurane (0)				
▸ nedocromil (2)				
▾ moxonidine (3)				
	Oxidation moxonidine → moxonidine.M1	0.48	271.66	+29.98
	Dehydrogenation moxonidine → moxonidine.M2	0.39	239.66	-2.02
	Hydroxylation moxonidine → moxonidine.M3	0.35	257.68	+16.00
▸ trimetrexate (3)				
▸ roxatidine (2)				
▸ chlormethine (1)				

Figure 44: CT-link Mlink spreadsheet.

And finally, in the MT-link spreadsheet tab, by clicking on a Molecule-Metabolite row, one can unfold the substantiation data generated by target profile comparison. Changes in the respective target profiles are colour-coded: green for overlapped targets, grey for substrate activities missing on metabolite and red for novel bioactivities present only in the metabolite profile (Figure 45).



Name	marketed_drugs_set
Status	finished

SUMMARY

Molecules	20 of 24
Metabolites	76

Molecule Identifier	Metabolite Identifier	Similarity
febuxostat	febuxostat.M1	
febuxostat	febuxostat.M2	
febuxostat	febuxostat.M3	
febuxostat	febuxostat.M4	
quinapril	quinapril.M1	
quinapril	quinapril.M2	
quinapril	quinapril.M3	
quinapril	quinapril.M4	
telbivudine	telbivudine.M1	
telbivudine	telbivudine.M2	
telbivudine	telbivudine.M3	
telbivudine	telbivudine.M4	
nedocromil	nedocromil.M1	
nedocromil	nedocromil.M2	
moxonidine	moxonidine.M1	
moxonidine	moxonidine.M2	
moxonidine	moxonidine.M3	
trimetrexate	trimetrexate.M1	
trimetrexate	trimetrexate.M2	
trimetrexate	trimetrexate.M3	
roxatidine	roxatidine.M1	
roxatidine	roxatidine.M2	
chlormethine	chlormethine.M1	
silodosin	silodosin.M1	
silodosin	silodosin.M2	
silodosin	silodosin.M3	
silodosin	silodosin.M4	
silodosin	silodosin.M5	
silodosin	silodosin.M6	
silodosin	silodosin.M7	
silodosin	silodosin.M8	
silodosin	silodosin.M9	
bazedoxifene	bazedoxifene.M1	
bazedoxifene	bazedoxifene.M2	
bazedoxifene	bazedoxifene.M3	
bazedoxifene	bazedoxifene.M4	
bazedoxifene	bazedoxifene.M5	
bazedoxifene	bazedoxifene.M6	

Figure 45: CT-link MTlink spreadsheet.

3.6.1.3.6 Selection, filtering and sorting

Selections are exclusively applied on entities: Molecules, Targets, Safety or Metabolism. In addition, selections can be only applied on a single entity at a time. In the following example (Figure 46), after selection of three molecules, all other links and entities are filtered accordingly. The number of molecules in the active selection is displayed in red background.

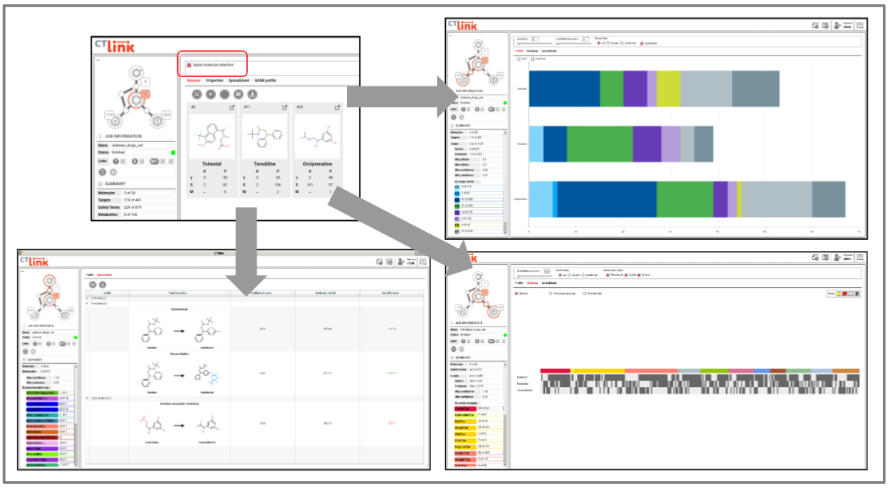
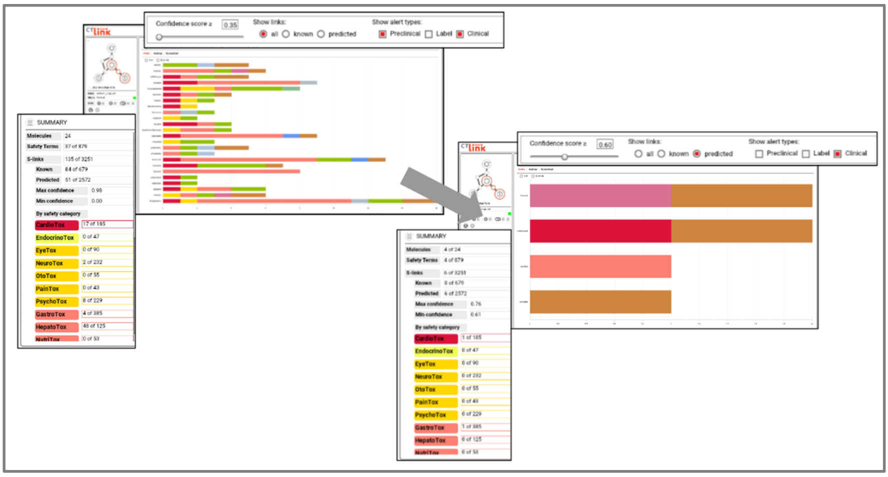


Figure 46: CT-link selection in action.

Filtering is exclusively applied on connections (T-link, S-link or M-link) and it can be simultaneously applied to any active entity selection. Any filtering readily affects the numbers in the *links* summary in the navigation panel, offering a highly interactive tool to the user (Figure 47).



Confidence score ≥ 0.35

Show links: ● all ○ known ○ predicted

Show alert types: ☒ Preclinical □ Label ☒ Clinical

SUMMARY

Molecules	24
Safety Terms	37 of 879
S-links	135 of 3251
Known	84 of 679
Predicted	51 of 2572
Max confidence	0.98
Min confidence	0.00
By safety category	
CardioTox	17 of 185
EndocrinoTox	0 of 47
EyeTox	0 of 90
NeuroTox	2 of 232
OtoTox	0 of 55
PainTox	0 of 43
PsychoTox	8 of 229
GastroTox	4 of 385
HepatoTox	48 of 125
NutriTox	0 of 53

Confidence score ≥ 0.60

Show links: ○ all ○ known ● predicted

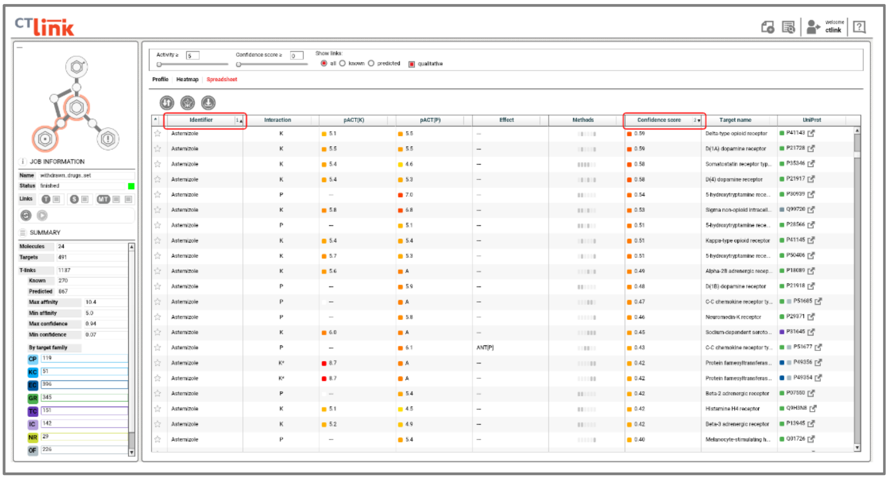
Show alert types: □ Preclinical □ Label ☒ Clinical

SUMMARY

Molecules	4 of 24
Safety Terms	4 of 879
S-links	6 of 3251
Known	0 of 679
Predicted	6 of 2572
Max confidence	0.76
Min confidence	0.61
By safety category	
CardioTox	1 of 185
EndocrinoTox	0 of 47
EyeTox	0 of 90
NeuroTox	0 of 232
OtoTox	0 of 55
PainTox	0 of 43
PsychoTox	0 of 229
GastroTox	1 of 385
HepatoTox	0 of 125
NutriTox	0 of 53

Figure 47: CT-link filtering in action.

All spreadsheets included in nodes and connections can be sorted by clicking on the column headers. Multiple sorting is possible. For example, Figure 48 shows the result of applying different filters: predicted (P) *links* are sorted, first, according to their molecular identifier and, second, according to the confidence score of the predictions.



JOB INFORMATION

Name	withdrawn_drugs_set
Status	finished

SUMMARY

Molecules	24
Targets	491
T-links	1137
Known	270
Predicted	867
Max affinity	10.4
Min affinity	5.0
Max confidence	0.94
Min confidence	0.07
By target family	
CP	119
KC	51
EC	396
GR	345
TC	151
IC	142
NR	29
OF	226

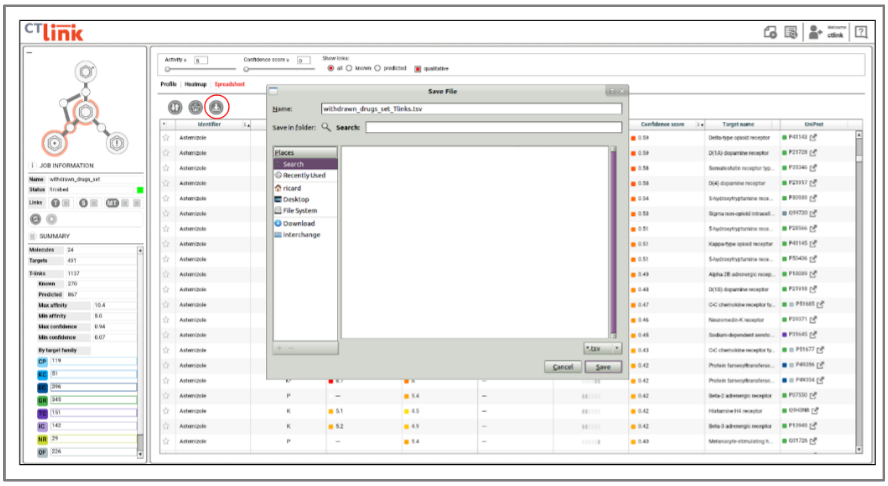
Activity ≥ 5 Confidence score ≥ 0

Show links: ● all ○ known ○ predicted ☒ qualitative

Identifier 1▲	Interaction	pACT(K)	pACT(P)	Effect	Methods	Confidence score 2▼	Target name	UniProt
Astemizole	K	5.1	5.6	—		0.59	Delta-type opioid receptor	P41143
Astemizole	K	5.5	5.5	—		0.59	D(1A) dopamine receptor	P21728
Astemizole	K	5.4	4.6	—		0.58	Somatostatin receptor typ...	P35346
Astemizole	K	5.4	5.3	—		0.58	D(4) dopamine receptor	P21917
Astemizole	P	—	7.0	—		0.54	5-hydroxytryptamine rece...	P30939
Astemizole	K	5.8	6.8	—		0.53	Sigma non-opioid intracell...	Q99720
Astemizole	P	—	5.1	—		0.51	5-hydroxytryptamine rece...	P28566
Astemizole	K	5.4	5.4	—		0.51	Kappa-type opioid receptor	P41145
Astemizole	K	5.7	5.3	—		0.51	5-hydroxytryptamine rece...	P50406
Astemizole	K	5.6	A	—		0.49	Alpha-2B adrenergic recep...	P18089
Astemizole	P	—	5.9	—		0.48	D(1B) dopamine receptor	P21918
Astemizole	P	—	A	—		0.47	C-C chemokine receptor ty...	P51685
Astemizole	P	—	5.8	—		0.46	Neuromedin-K receptor	P29371
Astemizole	K	6.0	A	—		0.45	Sodium-dependent seroto...	P31645
Astemizole	P	—	6.1	ANT[P]		0.43	C-C chemokine receptor ty...	P51677
Astemizole	K*	8.7	A	—		0.42	Protein farnesyltransferas...	P49356
Astemizole	K*	8.7	A	—		0.42	Protein farnesyltransferas...	P49354
Astemizole	P	—	5.4	—		0.42	Beta-2 adrenergic receptor	P07550
Astemizole	K	5.1	4.5	—		0.42	Histamine H4 receptor	Q9H3N8
Astemizole	K	5.2	4.9	—		0.42	Beta-3 adrenergic receptor	P13945
Astemizole	P	—	5.4	—		0.40	Melanocyte-stimulating h...	Q01726

Figure 48: CT-link multiple sorting.

There are different download buttons across the user interface. Those present in the links spreadsheets allow the user to download a tab-separated file (*.tsv) containing the visible data shown in the spreadsheet -selection, filtering and sorting included- for the active entity or connection (Figure 49).



Activity ≥ 5
Confidence score ≥ 0
Show links: ● all ○ known ○ predicted ☒ qualitative

Name: withdrawn_drugs_set_Tlinks.tsv
Save in folder: Search:

JOB INFORMATION

Name	withdrawn_drugs_set
Status	finished

SUMMARY

Molecules	24
Targets	491
T-links	1137
Known	270
Predicted	867
Max affinity	10.4
Min affinity	5.0
Max confidence	0.94
Min confidence	0.07
By target family	
CP	119
KC	51
EC	396
GR	345
TC	151
IC	142
NR	29
OF	226

Identifier	Confidence score	Target name	UniProt
Astemizole	0.59	Delta-type opioid receptor	P41143
Astemizole	0.59	D(1A) dopamine receptor	P21728
Astemizole	0.58	Somatostatin receptor typ...	P35346
Astemizole	0.58	D(4) dopamine receptor	P21917
Astemizole	0.54	5-hydroxytryptamine rece...	P30939
Astemizole	0.53	Sigma non-opioid intracell...	Q99720
Astemizole	0.51	5-hydroxytryptamine rece...	P28566
Astemizole	0.51	Kappa-type opioid receptor	P41145
Astemizole	0.51	5-hydroxytryptamine rece...	P50406
Astemizole	0.49	Alpha-2B adrenergic recep...	P18089
Astemizole	0.48	D(1B) dopamine receptor	P21918
Astemizole	0.47	C-C chemokine receptor ty...	P51685
Astemizole	0.46	Neuromedin-K receptor	P29371
Astemizole	0.45	Sodium-dependent seroto...	P31645
Astemizole	0.43	C-C chemokine receptor ty...	P51677
Astemizole	0.42	Protein farnesyltransferas...	P49356
Astemizole	0.42	Protein farnesyltransferas...	P49354
Astemizole	0.42	Beta-2 adrenergic receptor	P07550
Astemizole	0.42	Histamine H4 receptor	Q9H3N8
Astemizole	0.42	Beta-3 adrenergic receptor	P13945
Astemizole	0.40	Melanocyte-stimulating h...	Q01726

Figure 49: CT-link download.

Meanwhile, download buttons present in T-link and S-link substantiation panels or Metabolism browser allow the user to get a pdf report of the displayed information (Figure 50).

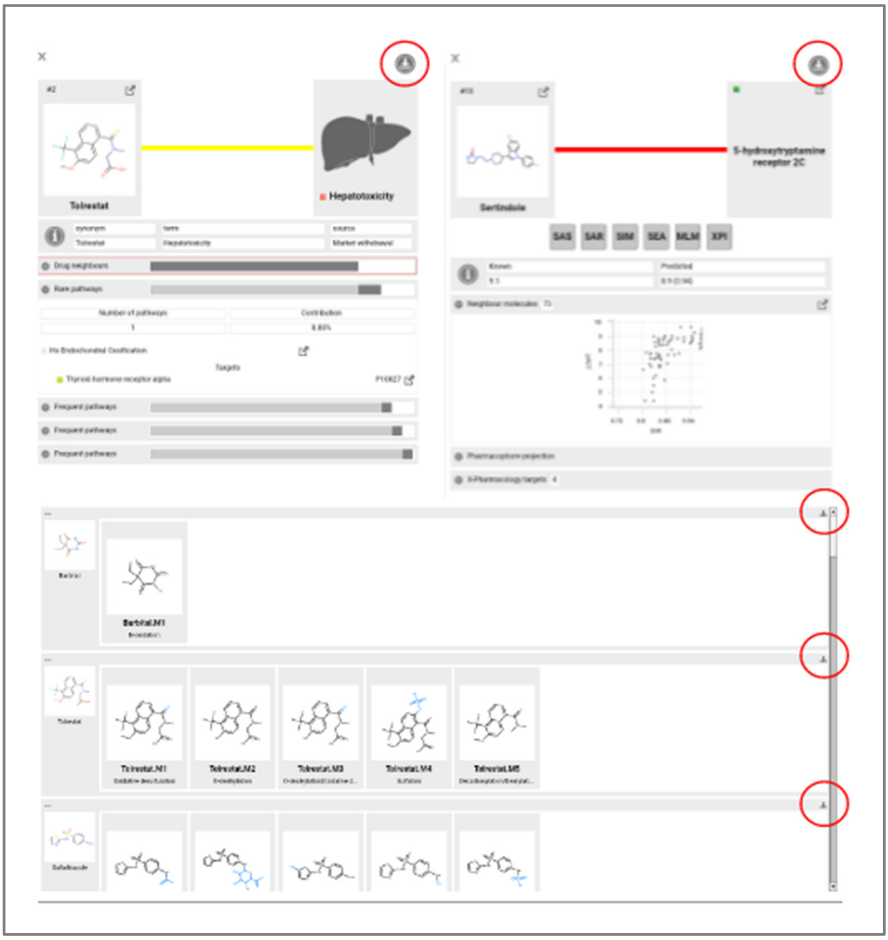


Figure 50: CT-link download substantiation.

A video with a demo of the CT-link platform is available here: https://www.youtube.com/watch?v=Va-zXa7KtC4

3.7 CLARITY

3.7.1 INTRODUCTION

After the release of the CT-link platform, we received feedback from users that identified some limitations of the platform and suggested some improvements to it. For example, the number of molecules that could be loaded in the platform was limited to 100 because of the huge amount of data and links associated with them and that had to be loaded in the computer of the user. One has to realize that the data is not only the structures of the 100 molecules but the millions of related data and links between them in the form of target interactions, safety alerts and generation of metabolites.

This amount of data also restricts the type of filters and interactions that can be offered to the user. Sophisticated filters, especially the ones that imply different entities and their relationships, are not possible with this architecture. Since the whole data is in the user computer, these actions cannot be done in an optimal way and we loose the responsiveness of the user interface. Clearly, this meant that the application had to be re-engineered to improve all aspects related to data management. In the user computer, we only need the data that is being visualized, not all the data. This is particularly true since often the data shown to the user is a summary or a subset of all the data generated.

In parallel, from the technical point of view, we found that almost all the browsers started to move away from Flash, many announcing that Flash will no longer be supported in the near future. Since the CT-link user interface was developed using Flex, which needs Flash to be renderered in the browser, that raised some concerns about sustainability and maintenance of the platform in the medium to long term. And finally, Adobe announced on July 2017 that it is planning to discontinue support for Flash by the end of 2020. These were the exact terms expressed by Adobe: “We will stop updating and distributing the Flash Player at the end of 2020 and encourage content creators to migrate any existing Flash content to these new open formats.”

With these new requirements and technical issues in mind, we decided to invest a huge amount of effort to develop a new platform called CLARITY that is going to be released on October 2017. It ought to be said that all previous clients of CT-link have been delighted with the beta experience of CLARITY and they will all transition to the new platform but also that CLARITY has generated a lot of interest among pharmaceutical and biotech companies, non-for-profit organizations, and academic institutions. CLARITY is the end of a long journey of developments in the field of data visualization tools for drug discovery and it represents a wonderful example of a successful technology transfer initiative. In the next sections, we are going to explain the technical details of this new platform in which we have been working for last year.

3.7.2 TECHNICAL RESULTS

In these sections, we are going to explain in more detail some important aspects of the CLARITY platform implementation that are not included in the manuscript.

3.7.2.1 FUNCTIONALITY

In this section, we are going to describe the functionality of the CLARITY platform:

- Create molecule sets: from uploading sdf/smi files or drawing molecules. Allows the user to add, modify or remove molecules from the set.
- Given a set of molecules, to explore known data
- Given a molecule set, to be able to execute:
  - Extract neighbours
  - Predict metabolites
  - Target profile, both known and predicted
  - Safety profile, both known and predicted
  - Target profile of metabolites, both known and predicted
- For each execution, the user can select the model against which predictions will be performed.

- Show molecules, metabolites, neighbours and jobs as cards, with a summary information for each entity.
- Get detailed dashboard information for a molecule, metabolites, neighbours and jobs.
- The user can apply filters, sort, and group all the information at any point of the application.
- Download the visualized information as a CSV.

3.7.2.2 DESIGN

The platform has been designed following the three-tier architecture model. The presentation layer controls the user interaction, the application layer deals with the logic of the application, and the model layer manages all data, both known and predicted (Figure 51).

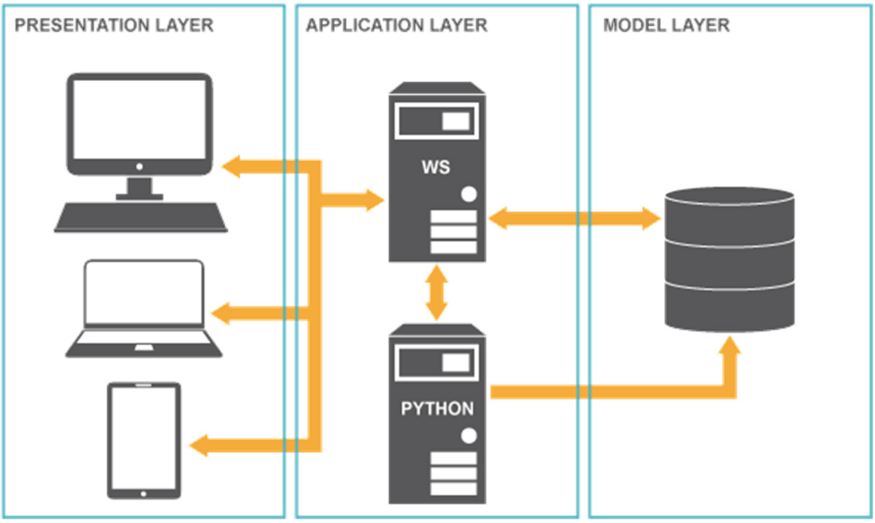


Figure 51: Clarity architecture.

3.7.2.2.1 Presentation layer

The presentation layer is designed following the Model-View-Control (MVC) pattern [42].

To develop a large-scale application using Angular, the idea is to start creating different components and then combine them together in an efficient way to build the whole application. Each component is an independent block with its own logic, view and data that is controlled, viewed and modelled.

In Angular, the MVC pattern is implemented using Html and TypeScript. The view is defined in HTML and the model and controller are defined in TypeScript.

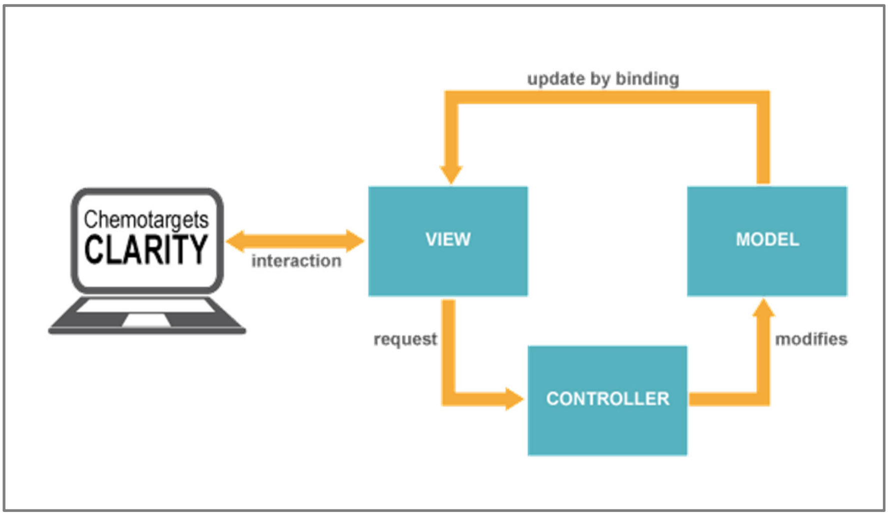


Figure 52: Clarity MVC model.

In Figure 52 we can see the MVC in action. These are the main steps followed:

- The user interacts with the application
- The controller calls the model to retrieve the information according to the user interaction.
- The controller receives the information and injects it to the viewer.
- The user receives the requested information.

3.7.2.2.2 Application layer

The application layer is divided in two interconnected modules:

- The prediction module
- The web services module

The prediction module generates the prediction data and it is beyond the scope of the present Thesis [35].

The web services module is developed using Scala programme language. The more important point in the application layer is the use of Slick [43] to generate dynamic SQL queries.

Slick ("Scala Language-Integrated Connection Kit") is a Typesafe's Functional Relational Mapping (FRM) [44] library for Scala that makes it easy to work with relational databases. There are two important concepts here: Typesafety and Functional Relational Mapping.

Let’s start with the latest concept, Functional Relational Mapping. It means that the database access is integrated inside the programming language used (in this case Scala) and we don’t write directly SQL queries. This approach allows you to work with database-stored data almost as if you were using Scala ordinary in memory data (such as collections) while at the same time giving you full control over database access and data transfer.

In fact, Functional Relational Mapping is a particular case of the Domain Specific Language (DSL) [45] concept. Slick is an embedded domain specific language for database access developed in Scala.

In addition, it is typesafe because the database access is pure Scala code so all the type checks and errors are detected in compilation time in contrast with the use of literal SQL queries, which are only checked at execution time. Moreover, because with Slick the database access code is pure Scala code and thanks to the functional programming paradigm, we can treat queries as functions that can be composed and enriched by using higher order functions like applying sorts, group and filters.

The language integrated query model in Slick's FRM is inspired by the LINQ project at Microsoft.

In our application, we use this approach to map the user interactions to the database queries. Specifically, we map all the entities and filters, sorts and group definitions from the user interaction to a database query that has to be executed.

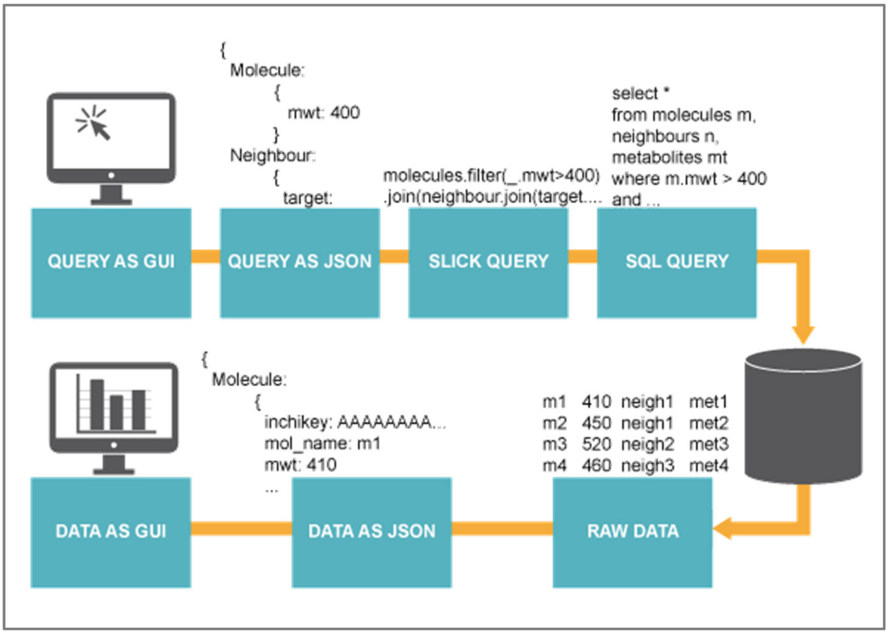
We are going to describe this functionality with an example.

Let’s assume the user wants to explore the results of a certain job. She/he wants to retrieve all the molecules that have a molecular weight greater than 400.

The user interaction is transformed into a query definition codified as JSON [46] and a request is sent to the web service.

The web service module receives the request, decodes the JSON query, transforms it into a Slick query, and the Slick query is executed.

The execution of the Slick query transforms the Slick query into a PostgreSQL query and the results are returned to Scala that codifies the results in JSON and returns it to the user view that ultimately shows the results to the user (Figure 53).



```
{
 Molecule:
        {
          mwt: 400
        }
 Neighbour:
        {
          target:
```

```
molecules.filter(_.mwt>400)
.join(neighbour.join(target....
```

```
select *
from molecules m,
neighbours n,
metabolites mt
where m.mwt > 400
and ...
```

```
m1  410  neigh1  met1
m2  450  neigh1  met2
m3  520  neigh2  met3
m4  460  neigh3  met4
```

```
{
 Molecule:
      {
        inchikey: AAAAAAAA...
        mol_name: m1
        mwt: 410
        ...
```

Figure 53: Query transformation.

This is a simple case with one entity (molecule) and a filter for this entity but ordinary research questions can be far more complex.

For example, imagine that the user wants to retrieve all the molecules that:

- have a molecular weight greater than 400
- with neighbours that have interaction with enzymes with an activity greater than 6
- with metabolites that have safety interactions with Hepatotox with a confidence score greater than 0.5

In this example, we are requesting for molecules but other entities are implied in the request: neighbours, targets, metabolites and safety. Each of these entities has its own filters. Just as the simple case, a query definition is generated and it is translated along the pipeline until the database.

The web service applies the filters to each entity and composes all the entities (Figure 54) to retrieve the molecules that pass the filters and the answer is returned to the viewer.

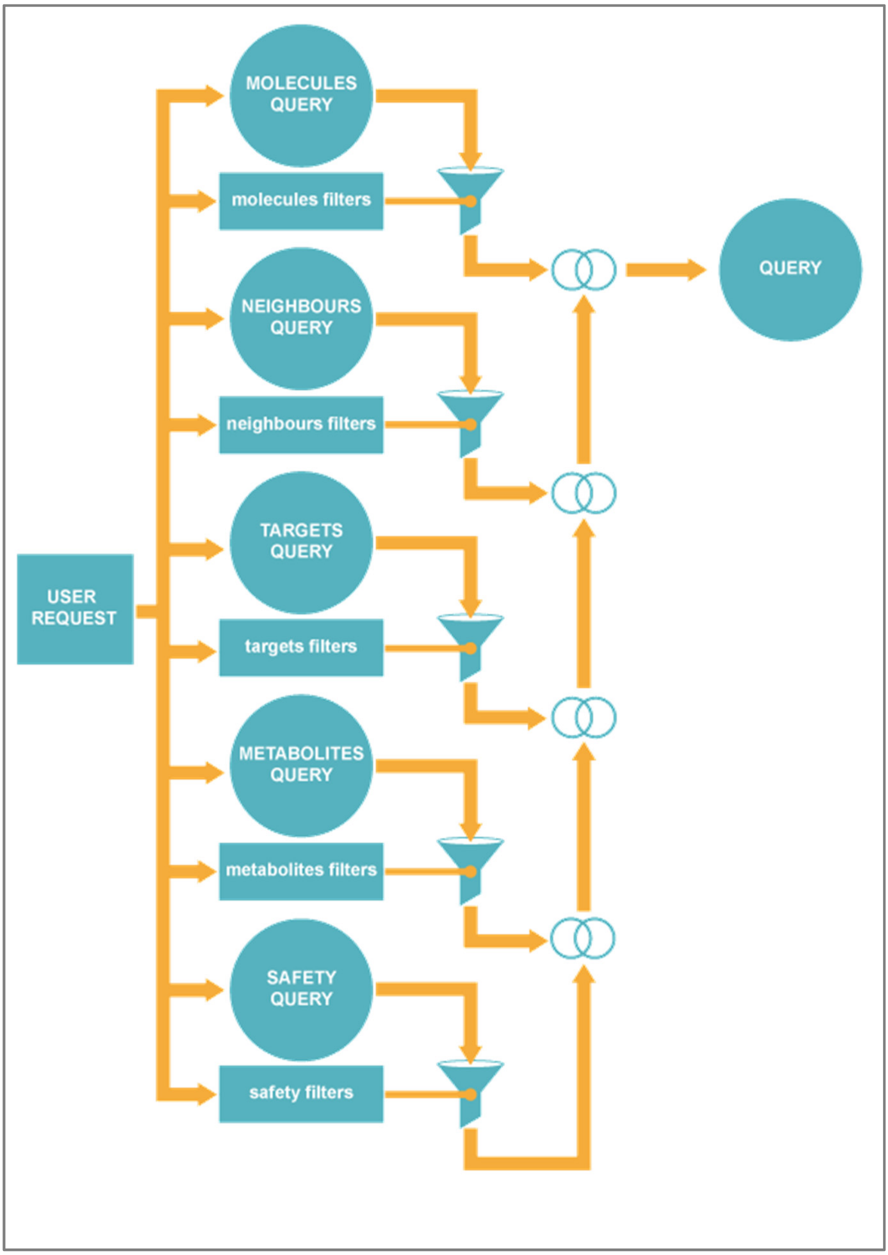


Figure 54: Filter application and query composition.

3.7.2.2.3 Data model

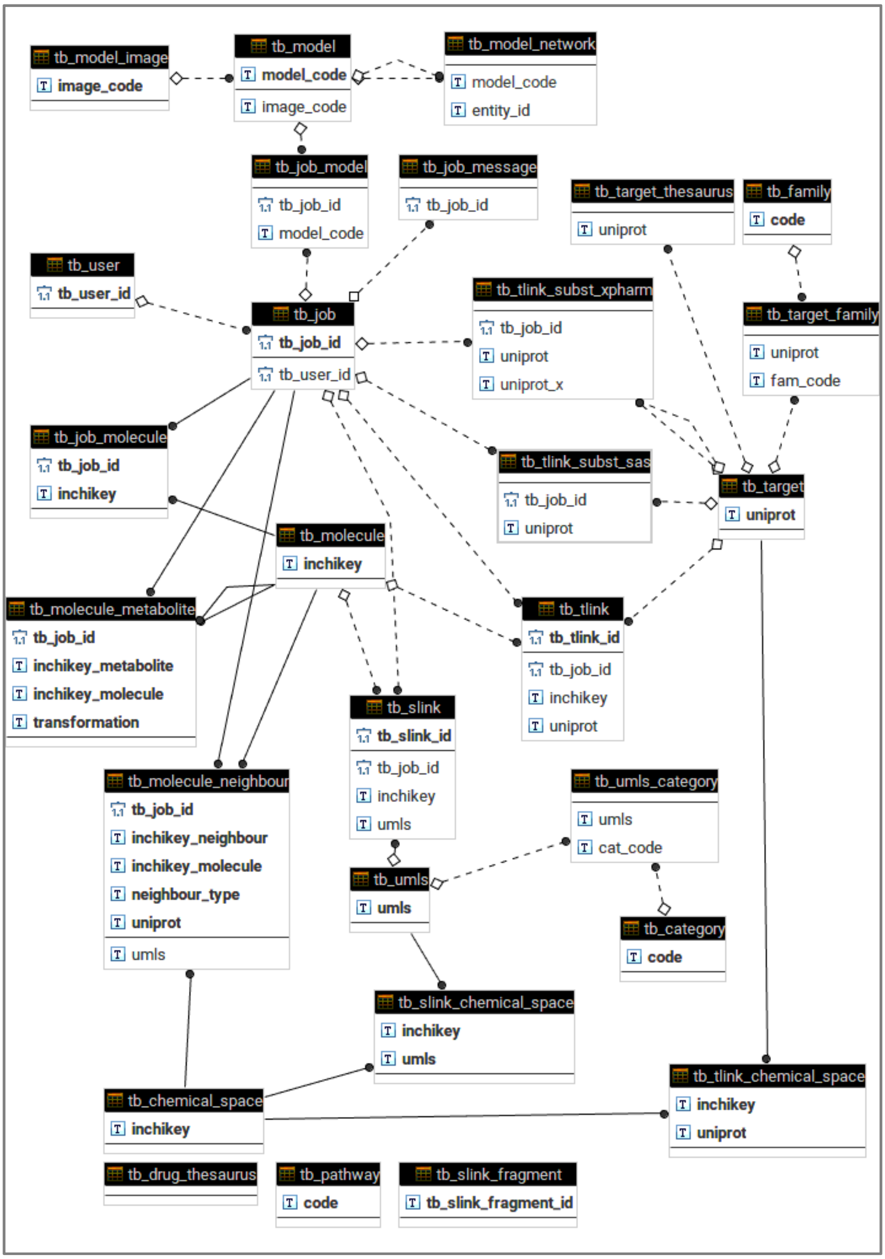


Figure 55: CLARITY data model.

The main tables of the model are:

tb_molecule	**contains information of the molecules stored in the platform.**
tb_molecule_metabolite	connects each molecule with its metabolites.
tb_molecule_neighbour	connects each molecule with its neighbours.
tb_tlink	connects a molecule with n targets according to a activity value known or predicted.
tb_slink	connects a molecule with n safety terns according to a safety association known or predicted
tb_chemical_space	contains all the molecules of the models used by the system

tb_tlink_chemical_space	contains tlink known data for all molecules in the models
tb_slink_chemical_space	contains slink known data for all molecules in the models

4 DISCUSSION

The amount of data relevant to drug discovery from all disciplines is increasing daily. These data are a very rich resource to extract new knowledge. But nowadays, the old ways to work with all these data have become outdated. This means that we have to move from plain files exploration to a graphical exploration of the data.

Graphical representation of the data is the best way to extract new knowledge from them. It offers a representation of very complex set of heterogeneous data in a simple way. With this simplification, we can be able to extract signals and detect patterns, which is what in the end drives the creation of new knowledge.

But just a graphical representation is not enough. A high level of interaction with this graphical representation is necessary. The user has to be able to interact with the graphs applying filters, sorting data and grouping it. Using these tools, the user can be able to make hypotheses, formulate new complex queries, and find answers that otherwise would have been impossible, or very difficult, to get.

5 CONCLUSIONS

The main contributions of this Thesis can be summarized in the following points:

- Implementation of the Pharmatrek web application as an exemplar of the Open PHACTS project. Pharmatrek provides to the user an intuitive, interactive and graphical way to explore the unified data provided by the OPS platform showing all the potency that this platform can provide.

- Implementation of the CT-link platform. In this platform, an existing predictive software (PredictFX) was combined with a new graphical user interface and a data storage system. We created a unified platform that allowed the user to execute predictions and navigate results in a graphical manner.

- Exceptional improvement of the CT-link platform into a new generation platform called CLARITY. This new platform solves some of the limitations we had in CT-link. With a modular architecture, it allows for updating the platform with new functionality without any impact to the client. In this way, we can be able to implement client suggestions as well as own new functionalities to the platform in a more flexible and efficient manner.

6 FUTURE WORK

The increase of data in the area of drug discovery is creating a complex network of different entities linked in various ways. The use of graphical tools to analyse and understand these relations is an evolving field and we expect to explore and test new methods. In addition, the inclusion of new entities implies new relations that may have to be explored with new graphical tools.

Another field to be explored is the ability of computers to analyze data using deep learning and automatic inference. These artificial intelligence methods can be applied to the extraction of knowledge from data.

Also, the interactive capabilities of the different devices are evolving very rapidly. These new interactive capabilities also provide new ways to interact with big heterogeneous data.

The field of data exploration is a continuously evolving field that requires adaptation to the new technological progress.

7 REFERENCES

[1] Digital Universe Invaded By Sensors. Available from: https://www.emc.com/about/news/press/2014/20140409-01.htm

[2] *Laney D.* Application Delivery Strategies. 2001; Available from: http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

[3] The 42 V's of Big Data and Data Science. Available from: http://www.elderresearch.com/company/blog/42-v-of-big-data

[4] *Manyika J (James), Chui M, Brown B, Bughin J, Dobbs R, Roxburgh C, Hung Byers A, McKinsey Global Institute.* Big Data : The next frontier for innovation, competition, and Productivity. McKinsey Global Institute, 2011 Available from: https://books.google.es/books/about/Big_Data.html?id=vN1CYAAACAAJ&redir_esc=y

[5] *Brin S, Page L.* The Anatomy of a Large-Scale Hypertextual Web Search Engine. Available from: http://infolab.stanford.edu/pub/papers/google.pdf

[6] Google AdSense. Available from: https://www.google.com/adsense

[7] Google AdWords. Available from: https://adwords.google.com/home/

[8] *Collaboration TA, Aad G, Abat E, Abdallah J, Abdelalim AA, et al.* The ATLAS Experiment at the CERN Large Hadron Collider. J Instrum 2008; 3: S08003–S08003 Available from: http://stacks.iop.org/1748-0221/3/i=08/a=S08003?key=crossref.b2ac868e8992413771c34191a8138368

[9] ROOT. Available from: https://root.cern.ch/

[10] Big PanDA. Available from: http://bigpanda.cern.ch

[11] *Aad G, Abajyan T, Abbott B, Abdallah J, Abdel Khalek S, et al.* Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC. Phys Lett B 2012; 716: 1–29 Available from: http://www.sciencedirect.com/science/article/pii/S037026931200857X

[12] LHC experiments delve deeper into precision | Media and Press Relations. Available from: https://press.cern/update/2017/07/lhc-experiments-delve-deeper-precision

[13] *Kuehn BM.* 1000 Genomes Project Promises Closer Look at Variation in Human Genome. JAMA 2008; 300:

2715 Available from: http://www.ncbi.nlm.nih.gov/pubmed/19088343

[14] *Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, Vaughan B, Preuss D, Leinonen R, Shumway M, Sherry S, Flicek P.* The 1000 Genomes Project: data management and community access. Nat Methods 2012; 9: 459–462 Available from: http://www.ncbi.nlm.nih.gov/pubmed/22543379

[15] *Zheng-Bradley X, Flicek P.* Applications of the 1000 Genomes Project resources. Brief Funct Genomics 2017; 16: 163–170 Available from: http://www.ncbi.nlm.nih.gov/pubmed/27436001

[16] *Anderson BC.* The End of Theory : The Data Deluge Makes the Scientific Method Obsolete The End of Theory : The Data Deluge Makes the Scientific Method Obsolete. 2008; 14–16 Available from: http://www.uvm.edu/~cmplxsys/wordpress/wp-content/uploads/reading-group/pdfs/2008/anderson2008.pdf

[17] *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

[18] *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://www.ncbi.nlm.nih.gov/pubmed/23548981

[19] *Allemang D, Hendler J.* Semantic Web for the Working Ontologist. 2011 Available from: http://linkinghub.elsevier.com/retrieve/pii/B9780123859655100044

[20] *Tukey JW.* Exploratory Data Analysis. Analysis 1977; 2: 688

[21] *Tukey JW.* Mathematics and the Picturing of Data*. 1974; Available from: http://www.mathunion.org/ICM/ICM1974.2/Main/icm1974.2.0523.0532.ocr.pdf

[22] *Fisherkeller MA, Friedman JH, Tukey JW.* Prim-9, an interactive multidimensional data display and analysis system. Dyn Graph Stat 1975; 91–109 Available from: http://www.slac.stanford.edu/pubs/slacpubs/1250/slac-pub-1408.pdf

[23] *Feynman RP.* The Theory of Positrons. Phys Rev 1949; 76: 749–759 Available from:

https://link.aps.org/doi/10.1103/PhysRev.76.749

[24] *Stolte C, Tang D, Hanrahan P.* Polaris: a system for query, analysis, and visualization of multidimensional relational databases. IEEE Trans Vis Comput Graph 2002; 8: 52–65 Available from: http://ieeexplore.ieee.org/document/981851/

[25] *Ahlberg C.* Spotfire: An Information Exploration Environment. SIGMOD Rec 1996; 25: 25–29 Available from: http://doi.acm.org/10.1145/245882.245893

[26] *Sander T, Freyss J, von Korff M, Rufener C.* DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. J Chem Inf Model 2015; 55: 460–473 Available from: http://pubs.acs.org/doi/abs/10.1021/ci500588j

[27] *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–D1107 Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkr777

[28] *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.ncbi.nlm.nih.gov/pubmed/22140110

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.ncbi.nlm.nih.gov/pubmed/24203711

30 *Pence HE, Williams A.* ChemSpider: An Online Chemical Information Resource. J Chem Educ 2010; 87: 1123–1124 Available from: http://pubs.acs.org/doi/abs/10.1021/ed100697w

31 *Garcia-Serna R, Ursu O, Oprea TI, Mestres J.* iPHACE: Integrative navigation in pharmacological space. Bioinformatics 2010; 26: 985–986 Available from: http://www.ncbi.nlm.nih.gov/pubmed/20156991

32 *Oliveira JL, Lopes P, Nunes T, Campos D, Boyer S, Ahlberg E, van Mulligen EM, Kors JA, Singh B, Furlong LI, Sanz F, Bauer-Mehren A, Carrascosa MC, Mestres J, Avillach P, Diallo G, D?az Acedo C, van der Lei J.* The EU-ADR Web Platform: delivering advanced pharmacovigilance tools. Pharmacoepidemiol Drug Saf

2013; 22: 459–467 Available from: http://www.ncbi.nlm.nih.gov/pubmed/23208789

[33] *Briansó F, Carrascosa MC, Oprea TI, Mestres J.* Cross-pharmacology analysis of G protein-coupled receptors. Curr Top Med Chem 2011; 11: 1956–1963 Available from: http://www.ncbi.nlm.nih.gov/pubmed/21851335

[34] Open PHACTS API. Available from: http://dev.openphacts.org

[35] *Garcia-Serna R, Mestres J.* Anticipating drug side effects by comparative pharmacology. Expert Opin Drug Metab Toxicol 2010; 6: 1253–1263 Available from: http://www.tandfonline.com/doi/full/10.1517/17425255.2010.509343

[36] *Vidal D, Mestres J.* In Silico Receptorome Screening of Antipsychotic Drugs. Mol Inform 2010; 29: 543–551 Available from: http://doi.wiley.com/10.1002/minf.201000055

[37] *Consortium W.* SPARQL W3 RECOMMENDATION. Available from: https://www.w3.org/TR/sparql11-overview/

[38] RDF. Available from: https://www.w3.org/RDF/

[39] RDF Schema. Available from: https://www.w3.org/TR/rdf-schema/

[40] *Odersky M, Spoon L, Venners B.* Programming in Scala: a comprehensive step-by-step guide. artima. 2010

[41] *Hilton P, Bakker E, Canedo F.* Play for Scala : covers Play 2. Manning, 2014

[42] *Krasner GE, Pope ST.* A Cookbook for Using the Model- View-Controller User Interface Paradigm in Smalltalk-80. Joop J Object Oriented Program 1988; 1: 26–49 Available from: https://www.lri.fr/~mbl/ENS/FONDIHM/2013/papers/Krasner-JOOP88.pdf

[43] *Typesafe.* Slick, functional relational mapping for Scala. Available from: http://slick.lightbend.com/

[44] *Wong L.* The functional guts of the Kleisli query system. Proc fifth ACM SIGPLAN Int Conf Funct Program - ICFP '00 2000; 35: 1–10 Available from: http://portal.acm.org/citation.cfm?doid=357766.351241

[45] *Mernik M, Heering J, Sloane AM.* When and how to develop domain-specific languages. ACM Comput Surv 2005; 37: 316–344 Available from: http://portal.acm.org/citation.cfm?doid=1118890.1118892

[46] The JSON Data Interchange Format COPYRIGHT

PROTECTED DOCUMENT. Available from: http://www.ecma-international.org/publications/files/ECMA-ST/ECMA-404.pdf

8 FIGURE INDEX