OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

# Deliverable Report D1.5

## Finalization of case studies and analysis of remaining weaknesses

This project is funded by

the European Union

OpenRiskNet: Open e-Infrastructure to Support Data Sharing, Knowledge Integration and *in silico* Analysis and Modelling in Risk Assessment

Project Number 731075

www.openrisknet.org

# Project identification

| | |
|---|---|
| **Grant Agreement** | 731075 |
| **Project Name** | OpenRiskNet: Open e-Infrastructure to Support Data Sharing, Knowledge Integration and *in silico* Analysis and Modelling in Risk Assessment |
| **Project Acronym** | OpenRiskNet |
| **Project Coordinator** | Edelweiss Connect GmbH |
| **Star date** | 1 December 2016 |
| **End date** | 30 November 2019 |
| **Duration** | 36 Months |
| **Project Partners** | P1 Edelweiss Connect GmbH Switzerland (EwC)<br>P2 Johannes Gutenberg-Universität Mainz, Germany (JGU)<br>P3 Fundacio Centre De Regulacio Genomica, Spain (CRG)<br>P4 Universiteit Maastricht, Netherlands (UM)<br>P5 The University Of Birmingham, United Kingdom (UoB)<br>P6 National Technical University Of Athens, Greece (NTUA)<br>P7 Fraunhofer Gesellschaft Zur Foerderung Der Angewandten Forschung E.V., Germany (Fraunhofer)<br>P8 Uppsala Universitet, Sweden (UU)<br>P9 Medizinische Universität Innsbruck, Austria (MUI)<br>P10 Informatics Matters Limited, United Kingdom (IM)<br>P11 Institut National De L'environnement Et Des Risques, France (INERIS)<br>P12 Vrije Universiteit Amsterdam, Netherlands (VU) |

# Deliverable Report identification

| | |
|---|---|
| **Document ID and title** | Deliverable 1.5 Finalization of case studies and analysis of remaining weaknesses |
| **Deliverable Type** | Other |
| **Dissemination Level** | Public (PU) |
| **Work Package** | WP1 |
| **Task(s)** | Task 1.3 |
| **Deliverable lead partner** | VU |
| **Author(s)** | Paul Jennings (VU), Philip Doganis, Pantelis Karatzas, Periklis Tsiros, Haralambos Sarimveis (NTUA), Lucian Farcal, Thomas Exner, Tomaz Mohoric (EwC), Atif Raza (JGU), Celine Brochot, Cleo Tebby (INERIS), Marvin Martens, Egon Willighagen, Danyel Jennen (UM) |
| **Status** | Final |
| **Version** | V1.0 |
| **Document history** | 2019-11-21 Draft version<br>2019-12-12 Final version |

# Table of Contents

# SUMMARY

The OpenRiskNet case studies (originally outlined in Deliverable 1.3) were developed to demonstrate the modularised application of interoperable and interlinked workflows. These workflows were designed to address specific aspects required to inform on the potential of a compound to be toxic to humans and to eventually perform a risk assessment analysis. While each case study targets a specific area including data collection, kinetics modelling, omics data and Quantitative Structure Activity Relationships (QSAR), together they address a more complete risk assessment framework. Additionally, the modules here are fine-tuned for the utilisation and application of new approach methodologies (NAMs) in order to accelerate the replacement of animals in risk assessment scenarios. These case studies guided the selection of data sources and tools for integration and acted as examples to demonstrate the OpenRiskNet achievements to improve the level of the corresponding APIs with respect to harmonisation of the API endpoints, service description and semantic annotation.

# INTRODUCTION

There are many challenges associated with accurately predicting human risk caused by chemical exposure. While whole animal studies have formed the basis of risk assessment in the past, this situation is changing with entire sectors transitioning to non-animal methods. The cosmetic industry have already moved away from the utilisation of animals for testing the safety of their chemicals. While non-human animal data is a useful basis for risk assessment, it is not always accurate for multiple reasons including; species differences in xenobiotic metabolism and transport, an overreliance on in-bred species and exaggerated dosing regimes. Also, utilisation of animal data does not address the fact that human beings are extremely varied in many ways including; genetics, lifestyle, geography and have complicated co-morbidities and co-exposures. A way forward is to focus on mechanistic aspects of chemical-induced toxicity and to build from this point adding in other aspects, layers and dimensions. A number of distinct but complementary sciences aim to facilitate a more modern approach to hazard identification and risk assessment. These can be loosely labelled as ***(1)*** mechanistic toxicology, ***(2)*** physiologically-based pharmacokinetic modelling and ***(3)*** cheminformatics and read across.

***1. Mechanistic toxicology***

Mechanistic toxicology aims to understand chemical interaction with biological systems and to elaborate on how such systems are perturbed by these interactions (if at all). The utility of omics technology and developments in systems biology is major driving force [1]. The utilisation of transcriptomics for example has provided a deeper understanding of how cells can adapt to perturbations by altering transcriptional programming to redress chemical induced homeostatic imbalances [2]. Transcriptionally orchestrated pathways, termed stress response pathways include Nrf2 which combats oxidative stress, p53 which combats genotoxicity and the unfolded protein response trio, ATF4, ATF6 and XBP1 combat proteotoxicity [3]. Transcriptomic experiments from carefully conducted *in vitro* studies are facilitating the deeper understanding of transcriptomic data from animal and clinical studies. Much of this data in the public domain is from publicly funded projects such as CarcinoGENOMICs, Predict-IV and TG-GATEs. Orchestration of this data into Adverse Outcome Pathways (AOP) facilitates the understanding of complex toxicology mechanisms

by the larger non-specialist community. AOPs aim to describe a specific adverse outcome, for example liver fibrosis, from its original initiation point (molecular initiating event, MIE), through a series of key events (KEs) that are causal, sequential and necessary for the disease outcome. While there are well described limitations to the AOP concept, the wide uptake and community based nature is facilitating the standardised utilisation of mechanistic toxicity throughout different fields of toxicology and the regulatory community [4].

***2. Physiologically-based pharmacokinetic (PBPK) modelling***

While hazard, the ability of a compound to initiate a biological perturbation, is a fundamental basis of toxicology, the concentration at a particular target is equally as important. Adsorption, distribution, metabolism and excretion processes dictate how much of a particular chemical or metabolite will reach a specific tissue in the body. PBPK modelling is a well-developed field and is based on computational compartmentalisation of anatomical, physiological, physical, and chemical descriptions of the phenomena involved. PBPK modelling can be used to translate animal dosimetry to other species including the human. Also such information can be used to predict *in vivo* doses from *in vitro* exposures by employing reverse dosimetry. This type of modelling is critical if *in vitro* based data is to be utilised in risk assessment regimes.

***3. Cheminformatics***

The ability to predict certain events from chemical properties is not a new concept and has been firmly embedded in the QSAR field. However, technical developments in computation combined with increasing knowledge of three dimensional native protein structure are further driving such activities. QSAR works best for direct ligand protein interactions, which is only a subset of toxicological interactions. This is further complicated by the ability of specific enzymes (e.g. cytochrome P450) to metabolise compounds. These events can also be predicted as the crystal structures of several xenobiotic metabolising enzymes have been resolved aiding the development of computational models. Read across has had large uptake in the risk assessment communities, where there is an attempt to garner toxicological information with data poor chemicals (target) from similarly structured chemicals where there is a lot of data (source). Read across was originally based solely on chemical structure similarity but is now beginning to use biological effect similarity, which aims to group compounds with specific biological activities regardless of chemical similarity [5].

There has been some efforts in previous EU projects, for example in the SEURAT and EU-ToxRisk projects to develop systems and strategies for chemical safety and risk assessment, by utilising existing animal data, conducting targeted human-based *in vitro* tests including toxicokinetics and toxicodynamic studies. One such proposal envisages the utilisation of only *in vitro* and *in silico* approaches [6]. This *ab initio* approach lays down a framework they propose could form the basis of a full risk assessment by guiding the evaluation through the different steps to be considered and enable and gain confidence in decision making.

The combination of mechanistic toxicology, PBPK with cheminformatic modelling is the basis for the development of modern risk assessment scenarios. Indeed, this forms the basis of the case study development in OpenRiskNet. The OpenRiskNet case study framework is composed of seven modules/cases. Case studies are already used in different research and innovation as well as infrastructure projects including SEURAT-1, EU-ToxRisk and NanoCommons to demonstrate that the objectives of the projects have been achieved and to obtain general acceptance. Shuttleworth defines a case study as “an in-depth study of a particular situation rather than a sweeping statistical survey. It is a method used to narrow down a very broad field of research into one easily researchable topic.

Whilst it will not answer a question completely, it will give some indications and allow further elaboration and hypothesis creation on a subject. The case study research design is also useful for testing whether scientific theories and models actually work in the real world."[7]

In the above mentioned projects, the case studies are organized around specific chemicals, mixtures or nanomaterials or groups of such. For these, risk assessment is performed including the experimental filling of data gaps if necessary either following the read-across or *ab initio* philosophy demonstrating that the newly developed tools are able to produce reliable results at least for specific toxicity endpoints or specific regulatory questions. In contrast, the goals of OpenRiskNet is not to establish new methods or risk assessment frameworks but to show how the integration of different existing tools into a common, harmonised platform gives users easier access to data and tools and showing them how combination of the tools can lead to higher confidence in the results. Each case study tackles a specific aspect required to facilitate risk assessment, but would not be enough on their own. However, the cases act in cooperation with the other modules to enhance the quality and depth of information required for hazard identification and risk assessment and the flexibility of the OpenRiskNet infrastructure allows for the re-use, combination and optimization to specific questions asked by academic and industrial researchers, risk assessors or regulators and, in this way, supports research projects, institutions, companies as well as individuals.

The **DataCure** module adds pre-existing information to the chemical ID and links to available toxicological and chemical databases. Mechanistic toxicology is addressed by the TGX, Sysgroup and AOPlink modules. **TGX** links the chemical to toxicogenomics-based data for prediction and mechanism identification. **SysGroup** provides integration of cheminformatic information with available omics data. **AOPLink**, facilitates the reannotation of information to be consistent with the developments in AOPs. PBPK is covered by the **RevK** module which provides access to PBPK models for forward and reverse dosimetry calculations. Cheminformatics and metabolism are addressed in ModelRX and MetaP, respectively. **ModelRX** creates a prediction modelling output based on QSAR. **MetaP** adds information about compound metabolism including predicting sites of metabolism.

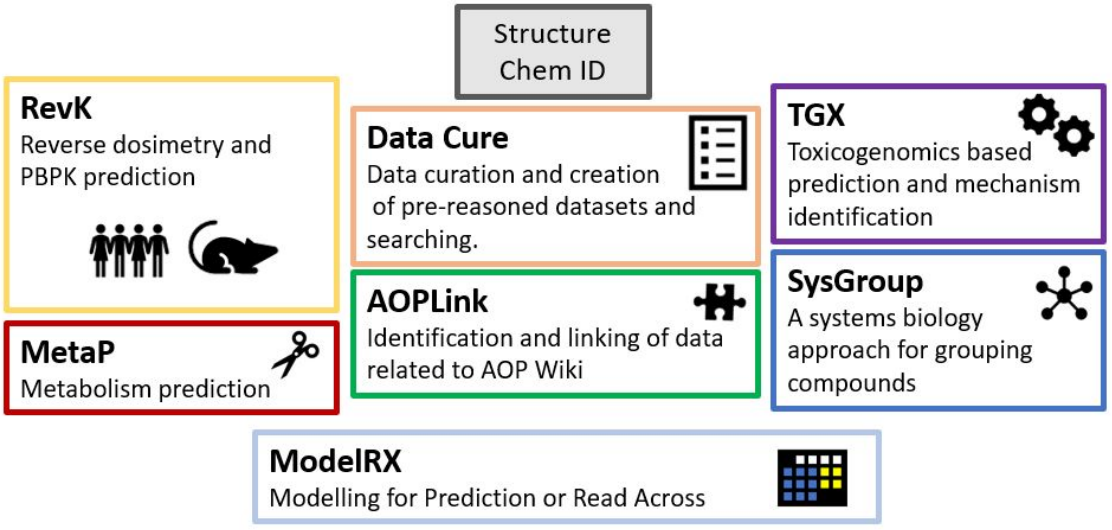



**Figure 1.** OpenRiskNet case study modules

# Relation to risk assessment frameworks

In the last decade toxicology and risk assessment have transitioned from animal-based histopathology to more a mechanistically driven data rich science, employing more *in vitro* assays and computational approaches. Modern risk assessment frameworks are no longer providing linear protocols of data generation and analysis but resemble more a decision tree, where data is coming from many areas, which are then combined and optimised for a specific problem at hand. This is prominently shown in the tiered framework proposed by Berggren et al. [7]. This proposed framework was used to define the OpenRiskNet case studies, where different modules of data collection or production using *in vitro* or *in silico* are executed and used for decision support according to the needs identified in earlier stages/tiers of the decision tree and transfer knowledge to the next stages. This is done until an assessment is possible and includes different exit routes, which can be taken when a threshold of toxicological concern (TTC) can be defined. Such modules include e.g. collection of support data, biokinetic refinement, point-of-departure (PoD) estimation and *in vitro - in vivo* extrapolation (IVIVE). A full list is given in **Table 1**. Each of these modules needs a specific workflow or set of workflows, from which the user can select, which must be flexible to facilitate interoperability with all the other modules in a non-predefined manner. This is done under the full control of the user so that each decision can be clearly followed based on the underlying evidence generated by the modules avoiding the risk that the process is becoming a black box. As an executable implementation of risk assessment frameworks like the one from Berggren et al. and others, integrated approaches to testing and assessment (IATA) and defined approaches try to formulate a preferred, somewhat standardised procedure but also acknowledge that specific parts might have to be adapted to the specific compound tested or other influence factors like availability of test methods also influencing the further processing, analysis and final model used in the assessment.

Other concepts and guidelines in the more specific area of using read-across for risk assessment of groups of chemicals also follow a similar modular approach. ECHA's Read-Across Assessment Framework (RAAF) clearly state the importance of risk estimation in contrast to hazard estimation [8]. Therefore, it has to be shown that the source compounds not only exhibit structural similarity to argue similar adverse effects and potency, but that they are also have similar toxicological properties, similar biokinetics properties and exposure scenarios, which corresponds to at least two different modules in the Berggren et al. [7] framework. Additionally, mechanistic understanding is essential for regulatory acceptance requiring additional data, tools and workflows. Based on the experience from case studies and mock submissions submitted to ECHA and evaluated by the regulatory advisory board, EU-ToxRisk is currently developing a platform providing general workflows for addressing all these different aspects of read across in a modular way, which can be selected and optimized for a specific problem at hand (*Manuel Pastor, private communication*).

**Table 1**. Workflow for the safety assessment of chemicals without using additional animal tests (adapted from Berggren et al. 2017 [6])

| Tier | Steps in the framework |
|---|---|
| Tier 0 - Identification of use scenario / chemical of concern / collection of existing information | 1. Identification of exposure / use scenario |
| | 2. Identification of molecular structure |
| | 3. Collection of support data |
| | 4. Identification of analogues / suitability assessment and existing data |
| Tier 1 - Hypothesis formulation | 5. Systemic bioavailability |
| | 6. Mode of Action hypothesis generation |
| Tier 2 - Application of the approach | 7A. Targeted testing |
| | 7B. Biokinetic refinement |
| | 8. Points of departure / IVIVE / uncertainty estimation / margin of safety |
| | 9. Final risk assessment or summary on insufficient information |

As described in **Deliverable D1.3** [9], the case studies of OpenRiskNet were designed to cover all important non-testing parts of the risk assessment framework, i.e. *in silico* modelling, simulations and predictions as well as data collection and interpretation. In the following description of each case study, we will use the modules defined in Berggren et al. [7], also used to classify the individual OpenRiskNet services in **Deliverable D4.3**, to show the relationships to this *ab initio* framework, i.e. risk assessment for compounds with no legacy animal data. This is not a limitation of the general applicability of OpenRiskNet results since the *ab initio* framework is a more detailed description of the general risk assessment framework outlined by the OECD. The latter can be roughly divided into four steps: hazard identification, hazard characterisation, exposure assessment and risk characterisation with clear, better defined counterparts in the *ab initio* framework. Also the different other guidelines for implementation of the OECD concepts, e.g. the SEURAT-1 concept [10], the ILSI-HESI RISK21 framework [11], a framework for the use of *in vitro*-derived biomarkers of toxicity in risk assessment [12] and a data-driven framework [13], propose similar modules, which will profit from OpenRiskNet implementations. In this way, OpenRiskNet case studies provide pieces of evidence, which can be put together to a full(er) risk assessment. However, a complete risk assessment for a given chemical was deliberately not attempted in OpenRiskNet as e-infrastructure projects generally do not create new data and without new data an assessment would soon be confronted with data gaps, which would render the results questionable. Instead, by linking different case studies, transferring data and knowledge from one to the other, reusing of workflows from multiple case studies, allows us to deal with increasingly more complex scenarios. We also provided guidance on how to utilise these solutions in single or multiple modules to support risk assessment by individual users or research projects.

More details are outlined in the case study descriptions and the discussion section below.

# CASE STUDIES DESCRIPTION

In this section, summaries of the case studies are provided to give the reader an overview of what was covered in the case studies, together with a self assessment of the results with respect to the main goals of OpenRiskNet including harmonization and interoperability, usability in risk assessment frameworks, readiness of the solution, possible generalisation to other tools and application scenarios, and remaining weaknesses. As described in the introduction, risk assessment frameworks require different types of information, which can be seen as independent modules. However, the output information then needs to be integrated for completion of a risk assessment. OpenRiskNet has used this modular approach which requires two levels of interoperability. First, the tools used inside one case study have to be able to talk to each other (intra-case-study interoperability). Secondly, the results produced have to be shared between modules/case studies (inter-case-study interoperability). Both aspects are considered here.

More details about the integrated tools, steps to combine and optimise them for the case studies and the obtained results can be found in the **Annex** and in a continuously updated version on the OpenRiskNet case studies website[1].

## DataCure - Data curation and creation of pre-reasoned datasets and searching

The detailed report in **Annex 1** and **online**[2].

### Description

DataCure establishes a process for data curation and annotation that makes use of APIs (eliminating the need for manual file sharing) and semantic annotations for a more systematic and reproducible data curation workflow. In this case study, users are provided with capabilities to allow access to different OpenRiskNet data sources and target specific entries in an automated fashion for the purpose of identifying data and metadata associated with a chemical or other endpoint of interest. The datasets can be curated using an OpenRiskNet services developed for this case study and re-submitted to the data source. Text mining facilities and workflows are also included for the purposes of data searching, extraction and annotation.

A first step in this process was to define APIs and provide the semantic annotation for selected databases (e.g. ToxCast/Tox21, TG-GATEs, DrugMatrix, FDA dataset, ChEMBL and ToxPlanet). These were then used to combine information from different resources, to provide them to automated curation processes for quality control and to make them fit for applications like further processing, QSAR modelling or mechanistic analysis, or, in the case of text resources, to make them available to automated text mining workflows to extract relevant information. In many cases, the obtained pre-reasoned and annotated datasets build intermediate results in an analysis or modelling workflow and will be consumed by other modules of the risk assessment process implemented in the other OpenRiskNet case studies. However, if considerable amounts of data or knowledge is derived from the original data or extracted and made available in a computer-accessible form relevant also for

[1] https://openrisknet.org/e-infrastructure/development/case-studies/

[2] https://openrisknet.org/e-infrastructure/development/case-studies/case-study-datacure/

other users, this data should be considered a new data resource and made available by OpenRiskNet as demonstrated with the corpus extracted by text mining or the processed TG-GATEs data.

## Self assessment

***How did the case study demonstrate the achievements of OpenRiskNet and how can this be generalized to other applications?***

Searching for and collecting of data is one of the most time consuming part in predictive toxicology, risk assessment and many other disciplines. Providing data via APIs allows for the automation of this process by developing cross-resource search and data access features. This was demonstrated with the workflows combining different primary sources like ToxCast and TG-GATEs. Additionally, curation workflows made interoperable with the data API can be used to improve the quality of the dataset by automatic validation e.g. of chemical identifiers and to extract information not until now available in a computer-readable form using text mining. All this curated data and information can then be fed back into the original data source to give other users access to this quality-checked information or can form completely new OpenRiskNet data and knowledge sources. As can be seen from the description above, data curation involves many different methods depending on the type of data and these had to be integrated and automated in different ways. Therefore, one could consider DataCure as a set of case studies in a case study.

***Which steps in the risk assessment framework does the case study support?***

This case study was clearly defined to support tier 0 (collection of existing information) and especially the steps 2 (identification of use scenario), 3 (collection of support data) and 4 (Identification of analogues/suitability assessment and existing data) of the risk assessment framework by generating structural identifiers from compound names, provide legacy support data and by identifying similar compounds using an advanced fragment-based approach. However, the text mining offerings also support step 1 by being able to search literature and reports for specific exposure scenarios. Since in the later tiers the newly generated data has to be combined with all the previous existing knowledge, workflows from DataCure are also beneficial here.

***Usability, Intra- and inter-case-study interoperability and automation***

Interoperability is achieved in OpenRiskNet based on the concept of harmonized application programming interfaces (APIs). Therefore, availability of APIs for the data sources was a prerequisite that data from these resources can be used in DataCure. For all but one of the resources mentioned above, OpenRiskNet partners provided the annotated APIs partly together with Python libraries for easier integration into the workflows. In this way, the processing and curation workflows could already benefit from the semantic annotation leading to a higher degree of automation of the data collection steps. The only exception is ChEMBL, where the APIs from the original data providers, which have not been harmonized and annotated following the OpenRiskNet concepts, were used. Even if the integration of the data needs more manual intervention of the user, which can still be performed by e.g. a Squonk workflow, the successful use of OpenRiskNet-compliant and third-party APIs shows that the OpenRiskNet approach for automated data access is already beneficial even if many additional data sources are not fully integrated yet.

Since DataCure needs to provide data for the subsequent multi-modular stages of risk assessment, a special focus was placed on inter-case study interoperability. The prepared processed, combined and/or pre-reasoned datasets must be made available to the consuming case studies in an easy accessible and harmonized way. This can be done by building a new data service with annotated APIs to access the generated data as it was done for the processed data of TG-GATEs. However,

such an approach is only sensible if the data is re-usable by many applications. For tasks performed in a specific risk assessment, a more flexible approach is needed since the data sources queried will be different from one study to the next. This was achieved, as in many other cases, by the use of partly predefined workflows. To demonstrate this, Squonk and Jupyter notebooks were developed to provide the input for the TGX, SysGroup, AOPLink and ModelRX case studies.

***Completeness / technical readiness***

This case study showed that the task needed for searching and collecting of data can be performed using the OpenRiskNet infrastructure and that it offers important tools for curating, quality checking and enriching of data. Especially the OpenRiskNet-compliant data sources have a high readiness and have been used in different applications inside OpenRiskNet (case studies, demonstrations) and are now transferred to other projects.

Data and the corresponding metadata, giving information on the resource but also the experimental protocol used to produce it, can be accessed via the APIs using standard functionality of many programming and scripting languages and from inside workflow managers like Jupyter and Squonk. To make access even easier, some of the APIs are complemented with programming libraries, which wrap the API calls into easy-to-use functions and provide optimised ways to create search queries. This allows to generate full automated workflows for data collection and curation. However, due to the diversity of the data source, this has to be made specific for one of them and combining of data from different sources needs the guiding hand of the user to specify what has to be combined. The semantic annotation is facilitating this by providing information, where specific data and metadata like compound identifiers is stored needed to combine the correct data. However, this annotation is only available for a limited amount of resources so far.

***Uptake***

As described above, workflows generated in DataCure were used in different other case studies for data access. Additionally, they were used in different training activities (e.g. hands-on sessions at the final OpenRiskNet workshop in Amsterdam). Specific services are now integrated into other projects like EU-ToxRisk to provide publicly available data but also for their internal data management. Workflows extending the demonstration workflows of DataCure are used in these projects for automated processing to e.g. generate benchmark doses for *in vitro* assays. Finally, well worth mentioning is the Implementation Challenge project with the Korean Institute of Toxicology (KIT), who now publicly provide their nanosafety dataset powered completely by OpenRiskNet solutions.

**Table 2**. DataCure relevant resources to promote uptake

| Title | Category | Related events |
|---|---|---|
| OpenRiskNet Part II: Predictive Toxicology based on Adverse Outcome Pathways and Biological Pathway Analysis | Poster | 2019-07-21 - ISMB/ECCB 2019 - International Conference on Intelligent Systems for Molecular Biology & European Conference on Computational Biology |
| Case Study description - Data curation and creation of pre-reasoned datasets and searching [DataCure] | Report | n/a |

| | | |
|---|---|---|
| Workflow: Access TG-GATEs data for selected compounds, select differentially expressed genes and identifier relevant pathways | Tutorial | n/a |
| Demonstration on data curation and creation of pre-reasoned datasets in the OpenRiskNet framework | Webinar recording and slides | 2019-03-18 - Demonstration on data curation and creation of pre-reasoned datasets in the OpenRiskNet framework |
| Workflows: practical example of Jupyter notebooks use, Data curation example, workflow across multiple case studies | Workshop session | 23 – 24 Oct 2019 / Amsterdam (NL), Final OpenRiskNet Workshop |
| DataCure flash presentation | Presentation | 23 – 24 Oct 2019 / Amsterdam (NL), Final OpenRiskNet Workshop |

***Analysis of remaining weaknesses***

OpenRiskNet was able to show the concept of annotated data APIs on a couple of examples data and knowledge sources. However, it is clear that this is only a small subset of all the available sources relevant to predictive toxicology. Even if we have been showing that also third-party resources, as long as they have an API, can be integrated, it is important that the approach is now taken up by many other players in the field to achieve the full benefits of OpenRiskNet. Collaborations with the associated partners, research and infrastructure projects like EU-ToxRisk and NanoCommons, and with large infrastructure initiatives like ELIXIR and EOSC have already shown that more resources can be expected in the near future.

Another remaining weakness was discovered during the annotation process of the OpenRiskNet data sources. The existing ontologies do not cover all requirements of the semantic interoperability layer. Therefore, some ontology development and design of the annotation process as an online or an offline/preprocessing step formed an ancillary part of this case study. This has to be continued as a community effort and, at least for the nanosafety area, the NanoSafety Cluster and NanoCommons are coordinating this.

# ModelRX - Modelling for Prediction or Read Across

The detailed report is available in **Annex 2** and **online**[3].

## Description

The ModelRX case study is concerned with providing predictive models based on user-provided data sets. As data can be generated from multiple sources and include values from experimental measurements, descriptors from computational methods and outputs from Omics-based analyses, the ModelRX case study was structured with flexibility regarding the nature of data to be processed and the inclusion of tools for the exploration of the chemical space with different approaches. A training data set is obtained from an OpenRiskNet data source, e.g. from DataCure, TGX, RevK or SysGroup case studies. Subsequently a model is trained with OpenRiskNet modelling tools and the resulting models are packaged into a container, documented and ontologically annotated. Model validation using OECD guidelines is provided. Once the modelling phase is finished, the users can use the created models for predictions on new chemicals.

The ModelRX case study allows previous analyses (i.e. from OpenRiskNet sources) to proceed into a modelling workflow that produces a model, allowing predictions regarding the (un)desired features of a compound thus guiding experimental work, reducing experiments (reducing or eliminating animal use) and leading to discoveries and improvements through a faster and more efficient path.

## Self assessment

***How did the case study demonstrate the achievements of OpenRiskNet and how can this be generalized to other applications?***

A vast number of *in-silico* predictive models have been developed and presented in the literature over the last decades. QSAR models are based on the concept of similarity and their usability and importance is continously increasing, because they can replace to a great extend animal testing and even costly *in vitro* experiments by predicting hazard related end-points for unknown chemicals using statistical/machine learning relationships which are built on known experimental data. However, in most situations, full QSAR models are provided in the literature only if they are relatively simple (like for example linear regression models with only a few independent variables) and even in this case, the models are not offered as software applications. For more complex QSAR models (such as neural networks) the model parameters are not given explicitly, and therefore the results cannot be reproduced and most importantly the intended use of the model is not satisfied, because interested stakeholders cannot actually apply the model to predict unknown toxicity end-points. OpenRiskNet has developed infrastructure for developing, collecting, sharing, testing and using QSAR models to compute predictions. All these functionalities are offered as open and easy to use web services, provided through Graphical User Interfaces and via harmonised APIs, which are interoperable with the data APIs. We demonstrated that multiple models produced by different partners can be interlinked to finally provide a consensus prediction. The modelling tools are now transferred to other projects and due to the technical flexibility and the microstructure architecture can easily be extended and generalised to other applications, even beyond the predictive toxicology discipline.

***Which steps in the risk assessment framework does the case study support?***

This case study was defined to mainly support tier 2 (hypothesis formulation) and especially the steps

[3] https://openrisknet.org/e-infrastructure/development/case-studies/case-study-modelrx/

7A (targeted testing) and 9 (final risk assessment) of the risk assessment framework by generating *in silico* predictive models, which can replace animal testing. *In silico* models can be used for hazard prediction and assessment, which are key information for defining points of departure and eventually performing risk assessment. The ModelRX case study also supports tier 0 because predictive models can guide the collection of the most relevant data and can assist in defining analogues based on statistical similarity methods.

***Usability, intra- and inter-case-study interoperability and automation***

We have developed tools that allow an inexperienced user to create, validate and finally expose his/her predictive model as a web application and make it available through both APIs and GUIs. Models are offered as free web applications which can be shared among groups, organisation or with the entire community. Models are accompanied with visualisation tools (for drawing a chemical or analysing the prediction results), detailed metadata (like for example descriptions of variables, ontological annotations), definitions of the domain of applicability or uncertainty, standard reports (QSAR Prediction Reporting Format, QPRF to verify compliance with OECD principles) and Predictive MarkUp Language (PMML) representations.

By structuring OpenRiskNet work and ModelRX in particular around harmonized APIs, there is a clear way for interoperability, both internally and externally with other case studies or applications. A number of Jupyter notebooks that demonstrate full use of modelling functionalities and interoperability of the ModelRX modelling tools has been made available on the designated Github space. We demonstrated that modelling workflows provided by different partners involved in ModelRX can be interlinked in a consensus modelling approach, allowing decisions to be made on the basis of an array of predictions from the various tools.

As modelling workflows start with data, datasets can be supplied by the user or originate from OpenRiskNet data sources, such as processed datasets in DataCure or datasets resulting from analysis performed in other case studies.

***Completeness / technical readiness***

The detailed description illustrates that the modelling infrastructure that has been developed and integrated in the context of the OpenRiskNet project achieved all the goals that were initially defined in this case study (developing, testing, validating, sharing, using models for predictions) and modelling services have achieved a high readiness level. Through APIs we have demonstrated that modelling components can be crosslinked to generate consensus modelling results and interlinked with other services providing input to risk assessment workflows.

***Uptake***

As described above, workflows generated in ModelRX were used in different other case studies for model building. Additionally, they received attention from a number of associated partners, who contributed by offering modelling services or actual models. More specifically, through the involvement of associated partners (NovaMechanics, BIGCHEM, Prosilico), we demonstrated how third-party modelling tools can be integrated within the OpenRiskNet infrastructure and be harmonised with the rest of the modelling services. We also highlighted that through Jaqpot, part of OpenRiskNet modelling infrastructure, we provide comprehensive and complete solutions to model developers, like KIT, that allow them to fully integrate their models and make them accessible to the community, as easy to use web services. ModelRX was demonstrated in webinars (one specifically for ModelRX) and in training activities (e.g. hands-on sessions at the final OpenRiskNet workshop in Amsterdam) and presented at conferences, with one poster focused jointly on ModelRX and RevK.

**Table 3**. Resources to promote ModelRX uptake (only those with a more intense focus on this CS are listed here)

| Title | Category | Related events |
|---|---|---|
| Demonstration on OpenRiskNet approach on modelling for prediction or read across (ModelRX case study) | Webinar recording and slides | 2019-06-11 - Demonstration on OpenRiskNet approach on modelling for prediction or read across (ModelRX case study) |
| Model RX OpenRiskNet - Case study using Jaqpot web modelling platform | Tutorial | n/a |
| Hands-on workshop session built around the ModelRX case study (support of the DataCure) | Workshop session | 23 – 24 Oct 2019 / Amsterdam (NL), Final OpenRiskNet Workshop |
| ModelRX flash presentation | Presentation | 23 – 24 Oct 2019 / Amsterdam (NL), Final OpenRiskNet Workshop |
| OpenRiskNet Part III: Modelling Services in Chemical/Nano-safety, Environmental Science and Pharmacokinetics | Poster | 2019-07-21 - ISMB/ECCB 2019 - International Conference on Intelligent Systems for Molecular Biology & European Conference on Computational Biology |
| Case Study description - Modelling for Prediction or Read Across [ModelRX] | Report | n/a |

***Analysis of remaining weaknesses***

The intense focus of OpenRiskNet to the risk assessment (RA) community, as exercised through the interactions with associated partners, other research projects and infrastructure initiatives like ELIXIR and EOSC, makes the results of ModelRX relevant and suitable for a wide audience that can take advantage of the open architecture, coverage of RA calculations and ability to interact with a multitude of external tools through APIs or Jupyter notebooks. We focused our development work on ensuring that modelling services are technically sound and can achieve a high level of interoperability. The services are not populated yet with a large number of predictive models, which means that we have not reached yet all the interested stakeholders, and this may have limited usability and uptake by the community. Hence, additional effort is needed to further disseminate the modelling services, so that OpenRiskNet infrastructure can play a key role in the predictive toxicology community and act as a central and harmonised repository for developing, sharing and actually using predictive models.

# SysGroup - A systems biology approach for grouping compounds

The detailed report in **Annex 3** and **online**[4].

## Description

This case study used the approach of the diXa / DECO2 (Cefic-LRI AIMT4) projects to reproduce and extend the results obtained on the identification of hepatotoxicant groups based on similarity in mechanisms of action (omics-based) and chemical structure using services from OpenRiskNet. We developed an initial workflow in which the different data types, i.e. structural similarity data, ligand-based target prediction data and transcriptomics data, are processed and prepared for data integration. The integration resulted in clusters of chemicals with similar properties based on the combination of the cheminformatic and omics data. However, due to time constraints we were not able to finalize the service and thus still some fine-tuning of the clustering is needed as well as curation of the clusters based on expert knowledge.

## Self assessment

***How did the case study demonstrate the achievements of OpenRiskNet and how can this be generalized to other applications?***

Chemical analogues identified by chemoinformatic tools are presumed to have a similar biological response and, consequently, have a relatively high likelihood to share similar toxicological properties. Therefore, the potential toxicity of target chemicals, for which no toxicity information is available, may be predicted by read-across from members of its chemical category (i.e. a group of chemical analogues with known toxicological properties). However, sometimes even seemingly subtle changes in chemical structure result in relevant changes in toxicological properties (for example, 2-AAF and 4-AAF, and APAP and AMAP. Therefore, it is difficult to predict with sufficient confidence the toxicological profile of a chemical solely on a chemoinformatic basis. This especially holds true for more complex toxicological endpoints related to repeated dose toxicity. Evaluation strategies based on integrating chemoinformatics approaches with mechanistic data from 'omics' have shown to strengthen confidence that the respective compounds within a class behave similarly in terms of their toxicological profile. This will allow for improved identification of toxicological hazards leading to better classification and labelling of chemicals.

OpenRiskNet provided the means to achieve this integration by creating a workflow partly consisting of services provided by OpenRiskNet. Currently, the workflow is already available for service developers via Gitlab.

***Which steps in the risk assessment framework does the case study support?***

SysGroup covers the identification of use scenario (step 1) / chemical of concern / collection of existing information (Tier 0 in the selected framework) and its steps related to:

- Identification of molecular structure (step 2);
- Collection of support data (step 3);
- Identification of analogues / suitability assessment and existing data (step 4).

[4] https://openrisknet.org/e-infrastructure/development/case-studies/case-study-sysgroup/

***Usability, Intra- and inter-case-study interoperability and automation***

For the developed workflow transcriptomics data was obtained from the preprocessed data made available through the TGX case study. With some minor adjustments also information and data from the EdelweissData service, developed under the DataCure case study, can be acquired. Ultimately, the resulting clusters from the workflow can feed into the case studies AOPLink and ModelRX.

Due to time constraints, the workflow is not finalized nor converted to a more comprehensible application such as Nextflow. Therefore, its usability and interoperability is lacking at this moment.

***Completeness / technical readiness***

Each step in the workflow is well documented in a Gitlab repository, but as indicated above, the workflow is not finalized yet.

***Uptake***

In its current state the SysGroup workflow is not yet ready as service for OpenRiskNet. However, the current workflow already contains tools that combined with the right existing services, can be turned into a service of its own. Additionally, individual steps are well documented and can be re-used by others building bioinformatics workflows to address their specific questions.

**Table 4**. SysGroup relevant resources to promote uptake

| Title | Category | Related events |
|---|---|---|
| OpenRiskNet Part II: Predictive Toxicology based on Adverse Outcome Pathways and Biological Pathway Analysis | Poster | 2019-07-21 - ISMB/ECCB 2019 - International Conference on Intelligent Systems for Molecular Biology & European Conference on Computational Biology |
| Case study description - A systems biology approach for grouping compounds [SysGroup] | Report | n/a |
| SysGroup flash presentation | Presentation | 23 – 24 Oct 2019 / Amsterdam (NL), Final OpenRiskNet Workshop |

***Analysis of remaining weaknesses***

Once the SysGroup workflow has been converted into a service of its own, the last step in the workflow, i.e. curation of the different clusters obtained, remains difficult to automate and requires expert knowledge to interpret its results. However, this is in line with the decision support/tree concept of modern risk assessment frameworks, were each module like the workflow implemented in SysGroup provides pieces of evidence to the risk assessor.

# MetaP - Metabolism Prediction

The detailed report in **Annex 4** and **online**[5].

## Description

Metabolites may well play an important role in adverse effects of parent drug (or other xenobiotic) compounds. In this case study VU (CS leader), HITeC/HHU (associate partner and implementation challenge winner), JGU and UU have worked together on making methods and tools available for metabolite and site-of-metabolism (SOM) prediction. For that purpose we have integrated and used ligand-based metabolite predictors (e.g. MetPred, FAME, SMARTCyp) and we incorporated protein-structure and -dynamics based approaches to predict the SOM by Cytochrome P450 enzymes (P450s). P450s metabolise ~75% of the currently marketed drugs and their active-site shape and plasticity often play an important role in determining the substrate's SOM. During method development, model calibration and validation we used databases such as XMetDB and other open-access databases for drugs, xenobiotics and their respective metabolites. To facilitate the combined use of the metabolite prediction approaches and their outcomes, we benefited of ongoing development in workflow management systems and we made Jupyter notebooks available to facilitate collection and visualization of results from the different available services. We illustrated the added value of having multiple predictors and our Jupyter notebooks available, in a pilot study on retrospective consensus predictions of known SOMs for drug compounds for which possible metabolite-associated toxicity was previously reported.

## Self assessment

***How did the case study demonstrate the achievements of OpenRiskNet and how can this be generalized to other applications?***

The successful integration in MetaP of various tools for SOM prediction into the OpenRiskNet platform and the subsequent seamless combination of their outputs into Jupyter notebooks is an excellent example of how a series of complementary tools can be combined and made available in a straightforward manner through the platform. As illustrated in this case study within a consensus study on SOM prediction, this can be of direct help for any aspect of risk assessment (or beyond) for which multiple individual prediction softwares are available, especially when these tools have been developed with different mechanistic backgrounds, training strategies and/or applicability domains in mind (and may thus well be of complementary and added value to each other).

***Which steps in the risk assessment framework does the case study support?***

The MetaP case study relates to Tier 0, step 1 (identification of molecular structure) and Tier 1, step 6 (mode of action).

***Usability, intra- and inter-case-study interoperability, and automation***

Users can directly employ the integrated tools and interpret their collective outcomes to decide on possible metabolite formation for a given (series of) query compound(s) of interest. The MetaP case study demonstrates service interoperability using Jupyter notebooks in which the integrated tools accept the same 3D structure format as input and return standardized SOM prediction data mapped

[5] https://openrisknet.org/e-infrastructure/development/case-studies/case-study-metap/

to the input atom IDs. Such interoperable data exchange facilitates aggregation of the output (e.g. in Pandas DataFrames) and visualization of the collective results (e.g. as 2D structure depiction using RDKit). MetaP predictions can be of great value in risk assessment within other case studies as performed during the project (and/or when using integrated workflows and tools), considering that metabolites often play a substantial role in possible adverse effects of parent compounds. It should be noted here that automated generation of molecular structures for potential metabolites is not available (yet), as our case study has primarily focused on SOM prediction, which is available in the majority of tools related to metabolite prediction.

***Completeness / technical readiness***

Seamless integration of services for SOM prediction was not only achieved within the timelines of the project by partners UU and VU, but also by associated partner and Implementation Challenge winner HITeC/HHU. As such, the MetaP case study has shown that the OpenRiskNet infrastructure is fully ready for uptake of (metabolism related) predictive tools and that their outputs can be directly combined and visualized using Jupyter notebooks. The palette of SOM and other predictors can be readily extended, as we plan to demonstrate e.g. by integration of modeling tools for microbial biotransformation in near future.

***Uptake***

As discussed in more detail in the **Annex**, the added value of having multiple complementary tools available for metabolite prediction was illustrated by the Jupyter notebook output for different molecules, which collects SOM predictions (and predictions of Phase I reaction types by one of the services). The compounds considered were selected based on reported toxicological effects related to their metabolism. Results of this consensus study and our experiences in MetaP service integration were presented late October 2019 in a plenary session during the final OpenRiskNet workshop in Amsterdam and during the most recent OpenTox meeting in Basel. We also plan to describe these findings in a manuscript to be submitted to the software section of a computational chemistry journal. In this manuscript we will also include results of and comparison with predictions by the ADME/PK predicting platform that was integrated into OpenRiskNet by associated partner and Implementation challenge winner *Prosilico*. In the meantime, interest in the MetaP case study and its outcomes has already been expressed by another H2020 project consortium (e.g. EU-ToxRisk).

**Table 5**. MetaP relevant resources to promote uptake

| Title | Category | Related events |
|---|---|---|
| MetaP flash presentation | Presentation | 23 – 24 Oct 2019 / Amsterdam (NL), Final OpenRiskNet Workshop |
| Site-of-metabolism prediction in OpenRiskNet | Presentation | 29 – 31 Oct 2019 / Basel, CH, OpenTox Euro 2019 Conference |
| Case Study description - Metabolism Prediction [MetaP] | Report | n/a |

***Analysis of remaining weaknesses***

The VU plasticity tool for Cytochrome P450 SOM prediction is currently the only OpenRiskNet service that needs 3D molecular structures as input. In this service, 2D to 3D structure conversion is handled

internally. Once other services become available that also use 3D structures as input, a dedicated OpenRiskNet service for 2D to 3D conversion will probably be needed such that structure conversion can be performed in a consistent way.

In addition and as noted above, the current case study output cannot directly (i.e. in a fully automated way) be used as input for other services, workflows or case studies, because at this moment the output does not comprise molecular structures of metabolites but SOMs (and reaction types) instead. In order to enable writing out metabolite structures by the services and/or Jupyter notebooks, rules have to be integrated that connect SOMs and reaction types to actual metabolites. In addition, more tools that consider reaction types may need to be integrated to allow for consensus prediction of this aspect as well. However, in its current status the MetaP and associated Jupyter notebook output is already of direct value to users as it aids in and greatly facilitates (consensus) predictions of possible metabolite formation.

# AOPLink - Identification and Linking of Data related to AOPWiki

The detailed report in **Annex 5** and **online**[6].

## Description

The Adverse Outcome Pathway (AOP) concept has been introduced to support risk assessment [14]. An AOP is initiated upon exposure to a stressor that causes a Molecular Initiating Event (MIE), followed by a series of Key Events (KEs) on increasing levels of biological organization. Eventually, the chain of KEs ends with the Adverse Outcome (AO), which describes the phenotypic outcome, disease, or the effect on the population.

In general, an AOP captures mechanistic knowledge of a sequence of toxicological responses after exposure to a stressor. While starting with molecular information, for example, the initial interaction of a chemical with a cell, the AOPs contain information of downstream responses of the tissue, organ, individual and population. Currently, AOPs are stored in the AOP-Wiki, a collaborative platform to exchange mechanistic toxicological knowledge as a part of the AOP-KB, an initiative by the OECD.

Normally, AOP development starts with a thorough literature search for existing knowledge, describing the sequence of KEs that form the AOP. However, the use of AOPs for regulatory purposes also requires detailed validation and linking to existing knowledge [15,16]. Part of the development of AOPs is the search for data that supports the occurrence and biological plausibility of KEs and their relationships (KERs). This type of data can be found in literature, and increasingly in public databases.

To support this data collection guided by information provided by the AOP-Wiki and replated knowledge bases like WikiPathways and AOP-DB, this case study established links between AOPs of the AOP-Wiki and experimental data based on stressors and target genes. This allowed finding AOPs related to experimental data, and finding data related to a particular AOP. Additionally, networks of AOPs were created automatically, which can be used to analyse synergistic effects of multiple stressors with different initiating event but the same adverse outcome or specific key events, which can result in different adverse outcomes.

## Self assessment

***How did the case study demonstrate the achievements of OpenRiskNet and how can this be generalized to other applications?***

The concept of AOPs has emerged and is increasingly popular to inform knowledge-based risk assessments. Their purpose is to capture mechanistic information of toxicological processes and structure in a way that is clear, pre-defined and modular. The majority of knowledge captured in AOPs is stored in the AOP-Wiki, and is based on scientific literature. However, as described above, AOPs and its building blocks require supporting data to verify the presence of the described biological processes, and the quantification of the relationships between sequential responses. Therefore, it is important to find ways of linking the quantitative descriptions of an AOP with experimental data and databases. Therefore, this case study showed how the AOP-Wiki and AOP-DB can be accessed through SPARQL endpoint and APIs how they can interoperate with other databases and repositories through the use of persistent identifiers and ontologies. Furthermore, BridgeDb has been integrated in

[6] https://openrisknet.org/e-infrastructure/development/case-studies/case-study-aoplink/

this case study to cover all the necessary mappings of identifiers between services and databases for chemicals, proteins, and more. In this way, information from the AOPs can now be used in Jupyter notebooks and seamlessly combined with other data and information coming e.g. for the SPARQL endpoint for WikiPathways, EdelweissData or the eNanoMapper database service to answer complex questions that required access to multiple services like data-driven Key Event and Key Event Relationship identification and validation of AOP outlines by experimental data of the stressors.

***Which steps in the risk assessment framework does the case study support?***

The main goal of the AOPLink case study was to provide information for given chemicals and nanomaterials that is captured in AOPs, and find supporting experimental data, which supports Tier 0 Steps 3 and 4. Furthermore, the knowledge captured in AOPs, which is explored in this case study, assists in Mode of Action hypothesis generation, which is Tier 1 Step 6 of the workflow for safety assessment of chemicals. This is one of the main aims the framework of AOPs, generalizing biological processes and capturing mechanistic knowledge from literature and experiments to facilitate the re-use of toxicological knowledge. That aspect, which is facilitated with the AOP-Wiki RDF development in AOPLink, allows hypothesis generation on the Mode of Action of chemicals of interest. The AOP-Wiki RDF also facilitates tasks related to Tier 3 Step 7A, by providing knowledge to initiate targeted testing for chemicals of interest, which is one of the main purposes of AOPs. Finally, AOPLink supports Tier 2 Step 9, by highlighting the sources of information (such as suitable assays), which helps to determine if an integrated approach to testing and assessment (IATA) can be developed based on the Key Events of the AOP and if this can be translated into a defined approach for the chemical(s) of interest. If this is not the case (yet), it can still be used to identify data and knowledge gaps.

***Usability, Intra- and inter-case-study interoperability and automation***

The central services of this case studies are knowledge repositories, being the AOP-Wiki, AOP-DB and WikiPathways. All three have been integrated in a similar fashion, by exposing RDF in a SPARQL endpoint, which allows for using SPARQL queries to extract parts of the data, through user interface or from a coding environment, such as Jupyter notebooks. The fact that they are all deployed as SPARQL endpoints, allows them to be used integratively by executing so-called federated SPARQL queries. However, time should be invested into increasing the usability of those services so that the users do not require to know the SPARQL query language and the RDF schema. Such work was initiated by using grlc (http://grlc.io/), which was explored during the project to build a user-friendly API.

More generally, a BridgeDb service has been integrated to allow mapping of identifiers between services, facilitating the interoperability of all tools, services and databases that work with chemicals, proteins, or gene identifiers, among others and forms an important link in the case studies but even more important to combine results from different case studies, which might have used different identifiers..

As all the tools in this case study are accessible through coding environments, such as Python or R, Jupyter notebooks allow the development of elaborate workflows that use the information captured in one of the SPARQL endpoints, and BridgeDb to facilitate linking of services by mapping identifiers. For example, as one of the main research questions in this case study is to find data to support a particular AOPs, a singular Jupyter notebook allows the combination of a range of tools from AOPLink and DataCure to answer that question. Whereas services from AOPLink provide knowledge and details about an AOP and its existing annotations, tools from DataCure are used to extract experimental data, which in turn, can be analysed by using the molecular pathways of WikiPathways.

***Completeness / technical readiness***

All services developed in this case study are ready to use through a variety of coding languages or risk assessment workflows in workflow systems, such as e.g. Jupyter notebooks. While the WikiPathways and AOP-Wiki SPARQL endpoints are complete and working, the AOP-DB SPARQL endpoint, which is the result of a successful implementation challenge, has only part of the data and will be further developed. The Virtuoso software used to host the SPARQL endpoints itself is an established, industrial product.

Regarding the completeness and readiness of the AOP-Wiki data, it should be noted that AOP-Wiki is not sufficiently machine readable to allow certain applications. For example, it is currently hard to link KEs to the genes or proteins involved, even if those are given in the textual description of the events. AOP-DB is filling this gap for existing AOPs but the information needs to be continuously manually curated to keep up with the ongoing AOP development. In parallel, efforts to directly annotate the AOP-Wiki with computer-readable metadata are ongoing.

***Uptake***

Because the idea of using AOPs in automated risk assessment workflows is relatively new, the AOP-Wiki RDF development was novel to the community. Therefore, the concept and RDF was promoted in a variety of occasions and was demonstrated briefly for small questions during meetings. One example is the Lorentz workshop on AOP e-resources, in which the AOP-Wiki SPARQL endpoint was used to extract all AOPs related to neuronal toxicity.
Furthermore, the US EPA has reached out to work with us after winning the OpenRiskNet Implementation Challenge, to initiate hosting AOP-DB data through the OpenRiskNet e-Infrastructure.

**Table 6**. AOPLink relevant resources to promote uptake

| Title | Category | Related events |
|---|---|---|
| OpenRiskNet Part II: Predictive Toxicology based on Adverse Outcome Pathways and Biological Pathway Analysis | Poster | 2019-07-21 - ISMB/ECCB 2019 - International Conference on Intelligent Systems for Molecular Biology & European Conference on Computational Biology, Switzerland |
| Enhancing the AOP-Wiki usability and accessibility with semantic web technologies | Poster | 7 – 11 Oct 2019 / Leiden, NL - Workshop ‘e-Resources to Revolutionise Toxicology: Linking Data to Decisions’, Leiden, Netherlands |
| Connecting Adverse Outcome Pathways, knowledge and data with AOPLink workflows | Webinar recordings and slides | 2019-07-15 - Connecting Adverse Outcome Pathways, knowledge and data with AOPLink workflows |
| Identification and linking of data related to AOPWiki | Webinar recordings and slides | 2019-03-26 - Identification and linking of data related to AOPWiki (an OpenRiskNet case study) |
| Case study description - Identification and Linking of Data related to AOPs of AOP-Wiki [AOPLink] | Report | n/a |

| | | |
|---|---|---|
| Remodelling of the AOP-Wiki and AOP-DB, and WikiPathways | Presentation | 2019-06-19 - Meeting of the Extended Advisory Group on Molecular Screening and Toxicogenomics (EAGMST), Paris, France |
| Workflows: practical example of Jupyter notebooks use, Data curation example, workflow across multiple case studies | Workshop session | 2019-10-23 - Final OpenRiskNet Workshop, Amsterdam, Netherlands |
| AOPLink flash presentation | Presentation | 23 – 24 Oct 2019 / Amsterdam (NL), Final OpenRiskNet Workshop |
| Discussions at the AOP-KB Face-to-face meeting | Meeting | 2019-11-12 - AOP-KB Face-to-face meeting, US EPA, North Carolina, USA |
| AOP-Wiki Resource Description Framework | Presentation | 2019-07-04 - Joint workshop NanoCommons - NanoSolveIT - RiskGONE |
| Expanding Adverse Outcome Pathway knowledge by creating AOP-Wiki RDF with semantic annotations to facilitate risk assessment of chemicals | Presentation and poster | 2 – 3 Apr 2019 / Lunteren, NL - BioSB - Bioinformatics & Systems Biology Conference 2019 |
| AOP-DB: The Adverse Outcome Pathway Database | Webinar recordings and slides | 2019-04-08 - AOP-DB: The Adverse Outcome Pathway Database |
| Introducing WikiPathways as a data-source to support Adverse Outcome Pathways for regulatory risk assessment of chemicals and nanomaterials | Peer-reviewed publication | n/a |
| Workflow: Access TG-GATEs data for selected compounds, select differentially expressed genes and identifier relevant pathways | Tutorial | n/a |

***Analysis of remaining weaknesses***

The AOPLink case study originally had two research questions to answer: find data to support AOPs, and to identify AOPs or KEs based on experimental data. While the developed and integrated tools assist in answering both questions, this case study was mostly focused on the first question. Therefore, we should further investigate the capability of identifying activated KEs based on experimental data of a chemical of interest. Approaches to generate such computationally-predicted AOPs (cpAOPs) were developed by OpenRiskNet partners and will now be implemented using AOPLink service.

Furthermore, the main ‘technologies’ that we used for the integration of resources and accessing data- and knowledge repositories were RDF and the SPARQL query language. Since this are still somewhat novel concepts in the toxicology area, most users would have to obtain new bioinformatics skills to use the tools. Because of that, there is a need for more APIs and User Interfaces to wrap around the SPARQL endpoint and allow all user communities to access and explore the data and knowledge contained in the RDF.

Finally, there are uncertainties about the AOP-Wiki, the central repository in this case study. Besides

the fact that most of the database has knowledge represented in free text fields and little use of consistent vocabularies or ontologies, most of the database lacks content and is incomplete. Therefore, the current possibilities of interoperability and linking of data with parts of the AOP-Wiki are limited. However, plans exist to include more types of predefined fields that use ontologies giving the many ongoing efforts the chance to add content to the repository in a more computer-readable manner.

# TGX - Toxicogenomics-based prediction and mechanism identification

The detailed report in **Annex 6** and **online**[7].

## Description

In this case study a transcriptomics-based hazard prediction model for identification of specific molecular initiating events (MIE) was applied based on top-down approaches. The MIEs could include, but are not limited to: (1) Genotoxicity (p53 activation), (2) Oxidative stress (Nrf2 activation), (3) Endoplasmic Reticulum Stress (unfolded protein response), (4) Dioxin-like activity (AhR receptor activation), (5) HIF1 alpha activation and (6) Nuclear receptor activation (e.g. for endocrine disruption).

We focussed on two top-down approaches for genotoxicity prediction. The first one, merely a proof of principle. resulted in the reproduction of an existing prediction model as a workflow initially using Snakemake workflow manager. Next this workflow was converted to a Nextflow-based workflow using OpenRiskNet's NextFlow service. The Nextflow version uses containerised steps, thus making it easier to deploy on any cloud infrastructure, and applicable to OpenRiskNet Virtual Environments. Finally, the Nextflow-based workflow was translated into a more generic approach so that it can be applied to other toxicogenomics studies.

The second top-down approach consisted of a metadata analysis for genotoxicity prediction. In this approach transcriptomic data on human, mouse and rat *in vitro* liver cell models exposed to hundreds of compounds were collected from the diXa data warehouse, NCBI GEO and EBI's ArrayExpress using the workflow from the first top-down approach. After preprocessing these data were made available as OpenRiskNet service.

## Self assessment

***How did the case study demonstrate the achievements of OpenRiskNet and how can this be generalized to other applications?***

In the past, several transcriptomics-based prediction models for genotoxicity or carcinogenicity have been developed and published. However, the repeatability and reproducibility of the models has not been shown and thus was addressed in this case study. As proof of principle, one of the prediction models from the Magkoufopoulou et al., 2012 article [17] was reproduced. It should be noted that the paper contained most of the information needed to reproduce the prediction model, but still didn't have 100% completeness. Having one of the authors in the implementation team of this case study was certainly helpful for correctly identifying the steps to follow and filling in some minor gaps. The developed workflow contains all the tools needed to perform the prediction analysis and together with full documentation including all parameters, repeatability and reproducibility are guaranteed.

Generalization of the data capture feature from the initial workflow is useful for developers to implement this into their own workflows. Additionally, by using Nextflow as the workflow manager, TGX can also be used as a showcase how OpenRiskNet services and workflows can profit from high-performance computer facilities and distributed cloud infrastructures.

Furthermore, the collected and processed transcriptomics data on multiple liver cell models from

[7] https://openrisknet.org/e-infrastructure/development/case-studies/case-study-tgx/

multiple species are directly available for further research by end-users, but also by developers that can incorporate the data as input into their own data analysis tools or workflows.

***Which steps in the risk assessment framework does the case study support?***

This case study is associated with all 3 tiers of the selected framework and in particular the following steps:

- Collection of support data (step 3);
- Identification of analogues / suitability assessment and existing data (step 4);
- Mode of Action hypothesis generation (step 6).

***Usability, Intra- and inter-case-study interoperability and automation***

The generic Nextflow-based workflow from the first top-down approach provides preprocessed external transcriptomic data from public resources on the public cloud from the VRE. This has been demonstrated for RNA sequencing data from NCBI.

The preprocessed transcriptomic data on human, mouse and rat *in vitro* liver cell models from the second top-down approach consists of average log2 ratios of each exposure (per time and per dosage) compared to its control. These data can directly be used by end-users to either investigate the gene expression changes of particular compounds or to use the data for prediction purposes. Furthermore, the data can feed directly into the case studies SysGroup to be used in the grouping of the chemicals, AOPLink to examine the pathways underlying a certain AOP, and ModelRX for prediction analyses.

In general, this case study demonstrates the usefulness and ease of use of the different workflows. Whereas the bioinformatics community already provide excellent interoperable tools and R libraries the workflow concept can now combine these and supply them in a predefined, repeatable and reproducible way to other users.

***Completeness / technical readiness***

The workflows, with each step well documented, are available in either a Gitlab or a Github repository. Furthermore, the preprocessed transcriptomics data are available as OpenRiskNet service and thus directly available for end-users and developers.

***Uptake***

Part of the preprocessed transcriptomics data was directly used in the SysGroup case study indicating its readiness to be used by others. Efforts to incorporate the data as well as metadata on the investigated chemicals into ToxicoDB have started and will continue beyond the end of the OpenRiskNet project.

**Table 7**. TGX relevant resources to promote uptake

| Title | Category | Related events |
|---|---|---|
| OpenRiskNet Part II: Predictive Toxicology based on Adverse Outcome Pathways and Biological Pathway Analysis | Poster | 2019-07-21 - ISMB/ECCB 2019 - International Conference on Intelligent Systems for Molecular Biology & European Conference on Computational Biology, Switzerland |

| | | |
|---|---|---|
| Use Nextflow for toxicogenomics-based prediction | Webinar recording and slides | 2019-05-27 - Use of Nextflow tool for toxicogenomics-based prediction and mechanism identification in OpenRiskNet e-infrastructure |
| Meta-analysis for genotoxicity prediction using data from multiple human in vitro cell models | Poster | 2018-09-02 - 54th Congress of the European Societies of Toxicology (EUROTOX) |
| Big Data in Toxicogenomics: Towards FAIR predictions | Presentation | 20-21 June 2018 - ICCA-LRI workshop: Demonstrating 21st Century Methods and Critical Tools for Risk-Based Decisions |
| Case Study description - Toxicogenomics-based prediction and mechanism identification [TGX] | Report | n/a |
| TGX flash presentation | Presentation | 23 – 24 Oct 2019 / Amsterdam (NL), Final OpenRiskNet Workshop |

***Analysis of remaining weaknesses***

Although data retrieval has been automated, manual curation was still needed, because of differences in the description of the datasets, e.g. differences in used ontologies, different metadata file formats.

# RevK - Reverse dosimetry and PBPK prediction

The detailed report in **Annex 7** and **online**[8].

## Description

This case-study demonstrates and documents the use of a web interface to physiologically-based pharmacokinetic (PBPK) models for forward and reverse dosimetry calculations. Forward calculations compute internal concentrations from given exposure doses. Reverse calculations compute exposure doses from internal concentrations or measured biomarker levels (e.g., urine concentration data). The result of those calculations can be used in risk assessments to help with *in vitro* to *in vivo* extrapolations or interspecies extrapolations.

The tools used for this case-study have been developed by NTUA re-using validated biokinetics models and are accessible through Jaqpot APIs and the Jaqpot user-friendly GUI for simulations. We have integrated the high throughput toxicokinetic (httk) R package, the PKSim software tool for whole-body physiologically based pharmacokinetic modeling, and we also provide means for developing and deploying user-defined models.

The case study is demonstrated through several reference chemicals or drugs: Imazalil for the httk model, Theophylline for the PKSim model, and Diazepam and Chlorpyrifos in rainbow trout for the user-defined models. The exposure scenarios chosen are in the range of corresponding environmental or therapeutic levels.

The steps required for developing PBPK models and deploying them through Jaqpot, performing simulations and obtaining and analysing the results are documented specifically for each tool workflow and in a detailed User Manual for the GUI.

## Self assessment

***How did the case study demonstrate the achievements of OpenRiskNet and how can this be generalized to other applications?***

Toxicokinetic modeling is an important step in chemical risk assessment since it provides a link between external exposure and internal concentrations, which are better predictors of toxicity. Physiologically-Based ToxicoKinetic (PBTK) models provide further refined predictions of internal kinetics. Model complexity, requirements regarding the amount of data necessary and lack of available tools hinder their use in risk assessment. Availability of models in an online tool can therefore facilitate and improve risk assessment.

PBTK models can be used to perform reverse dosimetry in order to estimate external exposure that leads to the levels observed in biomonitoring studies, using a given exposure scenario. The case study showed how reverse dosimetry can be performed in a user-friendly way with Jaqpot demonstrated with two substances with different PBTK models and providing realistic results. The upload of an existing generic PBTK model for fish had not been foreseen in OpenRiskNet proposal but now provides an example of how additional models can be added and shared.

***Which steps in the risk assessment framework does the case study support?***

The RevK case study relates to Tier 1.5 (biokinetics), and provides an application for Tiers 2.7B

[8] https://openrisknet.org/e-infrastructure/development/case-studies/case-study-revk/

(refined kinetics) and 2.8 (relating to IVIVE).

***Usability, intra- and inter-case-study interoperability and automation***

Users can use a generic PBTK model with parameters obtained from published databases or can upload their own PBTK model in the R language. Easy-to-use GUIs, available through the Jaqpot infrastructure, allow users to create dosing scenarios, run simulations, and generate concentration-time profiles in the form of tables or automatically generated figures. Tables with results can be downloaded as csv files for further processing and analysis. Depending on the complexity of the PBTK models, parameterisation can be time-consuming. We provide the option of using a .csv file rather than filling in individual input fields, which increases usability and traceability when the model in Jaqpot is run repeatedly with varying parameterizations. Through Jaqpot, RevK has integrated two open source popular pharma/toxico kinetic software packages, namely httk and PKSim. The RevK case study can be integrated with other case studies for exchanging of information and for building risk assessment workflows. In particular, partition coefficients, which are central in the development of a PBPK model can be estimated by predictive QSAR models developed in the ModelRX case study. Metabolic rates and pathways are also important components of biokinetics models, so interconnection with the MetaP case study can be very beneficial for refining metabolic rate estimations and for validating metabolism predictions.

***Completeness / technical readiness***

Jaqpot is running and operational. The output has been checked against results obtained directly with R. The RevK case study was carried out by two teams with complementary skills: NTUA was in charge of the technical part of designing and developing the tool and adding the existing models, Ineris was in charge or testing the interface and checking the predictions.

***Uptake***

BPP/TK models are very useful in risk assessment, since they can estimate internal exposure (concentration of chemical in different tissues and organs) as a result of an external exposure. Due to the complexity and the need to solve a system of differential equations, online web implementations of such models were lacking. Through the RevK case study, OpenRiskNet is closing this gap, as it has developed infrastructure for implementing and hosting practically any BPP/TK model as a web application, regardless of its complexity. The solution has already been adopted by other projects, like NanoCommons and has received the attention of the toxicokinetic community. The RevK functionalities were demonstrated and tested in various workshops and webinars during the lifetime of the project (for example Biokinetics training workshops in 1st Annual meeting in Basel, 2017 and in the OpenTox conference in Athens, 2018) and presented at conferences, with one poster focused jointly on ModelRX and RevK

**Table 8**. Resources to promote RevK uptake (only those with a more intense focus on this CS are listed here)

| Title | Category | Related events |
|---|---|---|
| RevK Pharmacokinetics OpenRiskNet Case study using Jaqpot web modelling | Tutorial (Video) | n/a |

| | | |
|---|---|---|
| platform | | |
| OpenRiskNet Part III: Modelling Services in Chemical/Nano-safety, Environmental Science and Pharmacokinetics | Poster | 2019-07-21 - ISMB/ECCB 2019 - International Conference on Intelligent Systems for Molecular Biology & European Conference on Computational Biology |
| Biokinetics training workshop | Workshop session | 21 – 23 Nov 2017 / Basel, CH - OpenTox Euro Conference |
| Biokinetics training workshop | Workshop session | 2018-10-10 - Hands-on Workshop on Biokinetics Modelling, 2018-10-10 - OpenTox EURO |
| Case Study description - Reverse dosimetry and PBPK prediction [RevK] | Report | n/a |
| Population pharmacokinetic reanalysis of a Diazepam PBPK model: a comparison of Stan and GNU MCSim | Peer-reviewed publication | n/a |

***Analysis of remaining weaknesses***

We have not demonstrated yet how the current output can be directly (i.e. in a fully automated way) be integrated with other services, workflows or case studies. This is to a large extent coupled with the nature of PBTK itself, where even for defining generic PBTK models on paper, they require a large number of parameters, many of which are often unknown and must be set to default values. This obstacle remains in our implementation of PBTK. Currently the software does not automatically use default values when input is missing. Additional models need to be implemented and be integrated into the OpenRiskNet infrastructure through the Jaqpot platform, to further attract the interest of the community and make OpenRiskNet a central repository for hosting, sharing and integrating biokinetics models into risk assessment workflows.

# CASE STUDIES DOCUMENTATION

The activities and resources related to the case studies were well documented and shared with case studies members and also with OpenRiskNet stakeholders. To achieve this, different repositories were established, e.g.:

- A dedicated **web page for each case study**: https://openrisknet.org/e-infrastructure/development/case-studies/. A detailed description of the case study was published including details on its definition, objectives, technical implementation, tools and services used and outcomes. Additionally, the list of services (automatically imported from the Services Catalogue) and related resources (automatically imported from the project Library) are directly accessible (see a representative example shown in **Figure 2**).

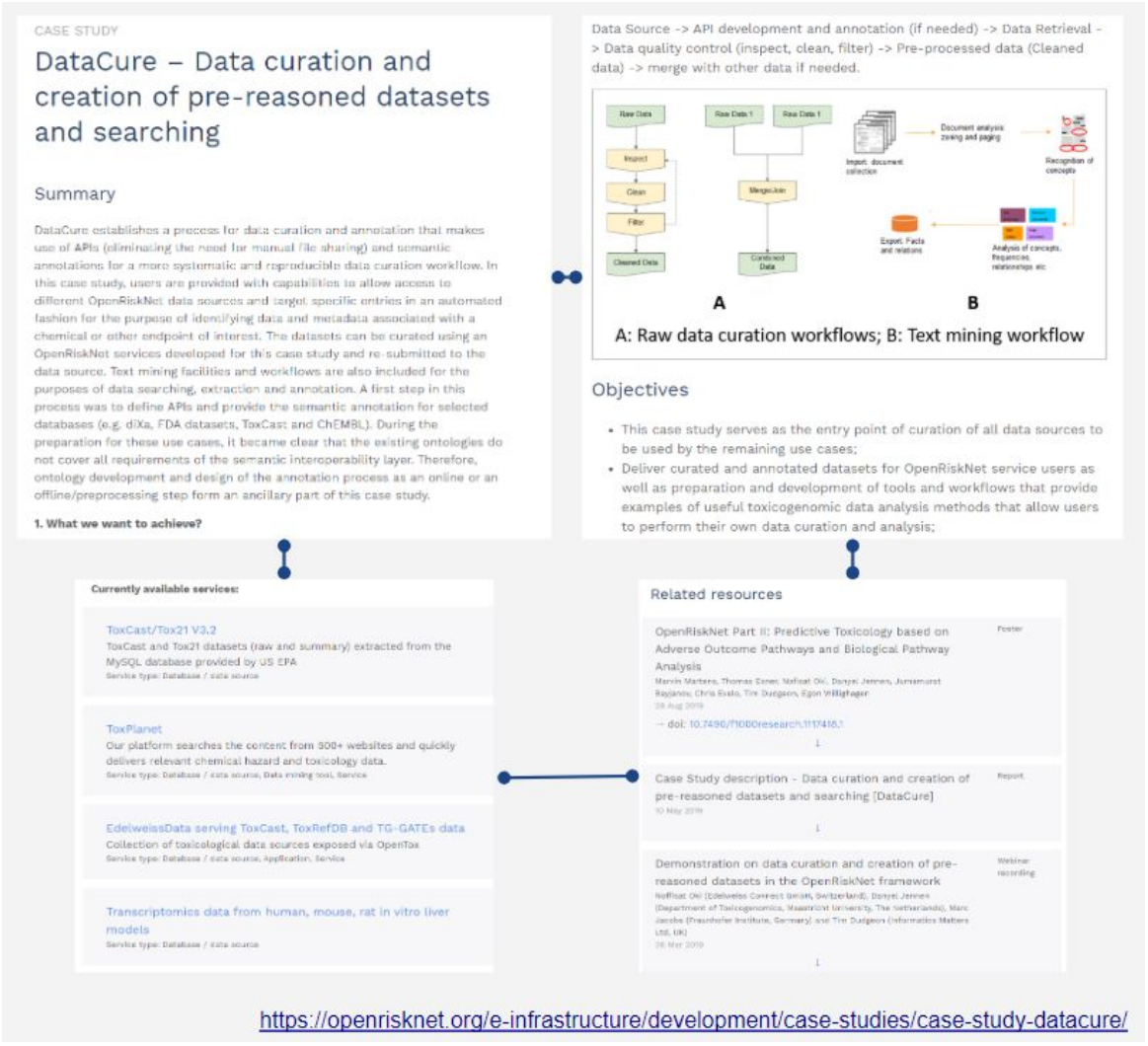



**Figure 2**. Example of documentation of DataCure case study on the OpenRiskNet website

- **Repository of codes and notebooks** - the workflows related to the case studies were developed and published in GitHub online and open access repository: https://github.com/OpenRiskNet/notebooks. This online repository was used by project members to collaboratively develop different workflows that can be used in Jupyter and Squonk notebooks by any user in order to integrate and access OpenRiskNet services (**Figure 3**).

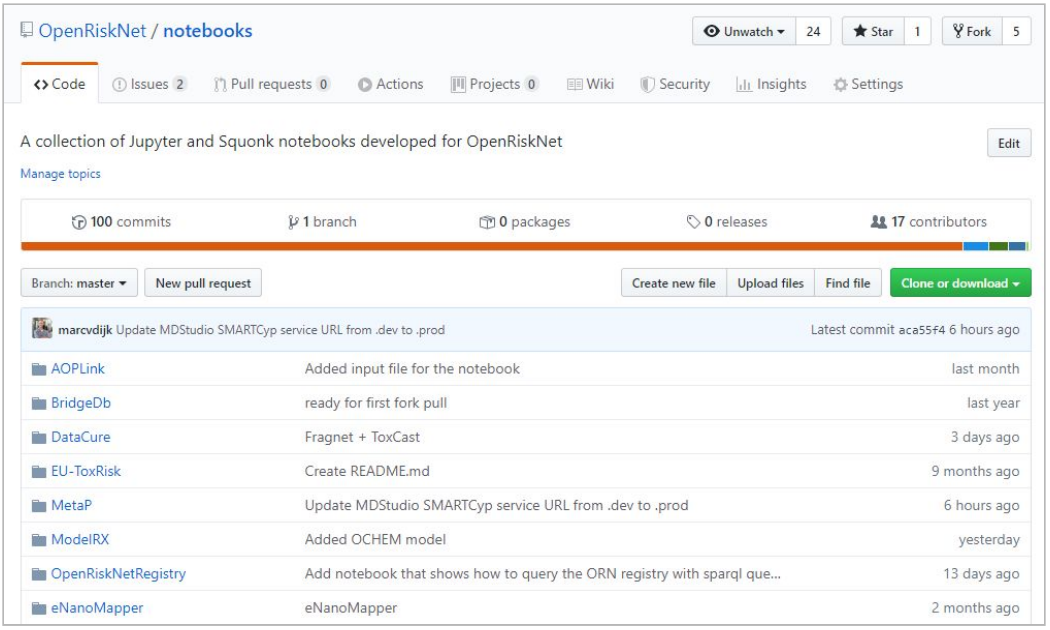



**Figure 3**. Screenshot from the GitHub repository dedicated to notebooks

- **Case study shared documents**: the contribution of multiple members was facilitated by using shared documents for each case study, including here a main file ("case study description") used further as a **case study report**.

**Demonstration sessions** (online or workshops) were organised to complete the collection of support materials on case studies implemented. Users were invited to become familiar with the OpenRiskNet concepts before using its services. A series of **webinars**[9] were organised on demonstrating these capabilities and allowing users to interact with OpenRiskNet service providers. Webinars addressed specific topics related to the case studies, e.g.:

- Demonstration on data curation and creation of pre-reasoned datasets in the OpenRiskNet framework (18 March 2019);
- Identification and linking of data related to AOPWiki (an OpenRiskNet case study) (26 March 2019);
- AOP-DB: The Adverse Outcome Pathway Database (8 April 2019);
- Use of Nextflow tool for toxicogenomics-based prediction and mechanism identification in OpenRiskNet e-infrastructure (27 May 2019);
- Demonstration on OpenRiskNet approach on modelling for prediction or read across (ModelRX case study) (11 June 2019);
- Connecting Adverse Outcome Pathways, knowledge and data with AOPLink workflows (15 July 2019).

Sessions on case studies were also included in face-to-face events organised by OpenRiskNet

[9] https://openrisknet.org/library/?category=Webinar+recording&audience=&organisation=

members. For example, a special section of the final workshop was dedicated to the case studies presentation and practical demonstrations. Details on these activities are included in **Deliverable 3.5**.

All materials including the recordings of the webinars, the slides of the presentations and the supporting materials in the form of tutorials, wiki pages and downloadable versions of the executed workflows are available from the OpenRiskNet webpage[10]. Alternatively, the recordings can be accessed together in the OpenRiskNet YouTube playlist[11].

[10] https://openrisknet.org/library/
[11] https://www.youtube.com/playlist?list=PLTxtsS5QQK1ymTJPa2aTIGfUNLt-3M4I0

# DISCUSSION AND CONCLUSION

The OpenRiskNet case studies have been developed to allow automation and functional integration for applications in hazard identification and risk assessment. Each of the seven case studies address specific areas of the risk assessment process where automation is highly desirable. The case study workflows and the services used therein have been deployed, are functionally operational with proven interoperationality and can now be re-used and adopted as modules of risk assessment frameworks according to the different steps defined by Berggren et al. [6] and other similar frameworks. Usefulness of the OpenRiskNet approaches and especially the interoperability of the provided service could already be proven by the workflows developed for these parts/modules and on a higher level by the programmatic linking of case studies like DataCure and AOPLink to just name one example.

Services from third parties included as part of the Implementation Challenge were able to complement the coverage of methods needed in the case studies and demonstrated the interoperability and combinability of services beyond the OpenRiskNet consortium. Integration of third-party tools in the MetaP, AOPLink and ModelRX studies especially strengthened the consensus approaches demonstrating the benefits of access to multiple models with different strengths and weaknesses and even provided the missing links between the AOP knowledge sources and the experimental data. Thus, the most important project objectives i.e. to generate prototypes for service integration, provide best-practice examples and to show the usefulness and transferability of the concepts was successfully achieved and demonstrated in functional integration workflows. As outlined in the Sustainability Plan, the OpenRiskNet consortium with its associated partners will continue to provide their services on OpenRiskNet. Also we aim to further optimise and develop these tools by getting involved in further infrastructure and research projects in the field of toxicology and risk assessment. Additionally, collaborations have been established to linking and aligning with other e-infrastructures of neighboring disciplines and to continuously increasing amount of commercial services.

A demonstration of the successful uptake of OpenRiskNet modules is the collaboration with the European Joint Programme on Rare Diseases (EJP RD). In an implementation challenge led by EJP RD partner LUMC, we developed a new “link set” that uses information from the comparative toxicogenomics database (CTD) to link toxic compounds to their protein targets. This link set can be used by the CyTargetLinker plugin for the network biology tool Cytoscape to evaluate which toxic compounds hit a rare disease network all using the OpenRiskNet concepts of workflows e.g. distributed as Jupyter notebooks. A more advanced approach currently implemented by the EJP RD will use molecular versions of adverse outcome pathways to create molecular networks in the way demonstrated in the AOPLink case study. Comparison of overlap and mutual influences between the rare disease and adverse outcome pathway networks can then be used to judge potential affected risks for classes of compounds that share the same AOP.

The next logical step will be to attempt a full risk assessment on a specific set of compounds by using the workflows developed in the case studies. Such an exercise would need some adoption and optimisation work. Optimisation would be required both on the data and tool side. From the data side a selection of the most relevant chemicals for risk assessment, and collection of sufficient data and data types would likely require the commissioning of new *in vitro* and *in silico* experiments to fill data gaps and to decrease uncertainty. From the tool side, we would likely need to amend some of the modules with additional services to cover additional toxicology endpoints not addressed by the case studies as well as to give the user more options to find the optimal data and tools and combination thereof to make the best informed decisions throughout the different risk assessment tiers. This will, as described above, come from the ongoing sustainability efforts of the consortium and increasing network of associated partners. This will be supported beyond the lifespan of the project by bringinging OpenRiskNet concepts, workflows and services in contact with other consortia that are

developing tools and produce new data in a more focused way to support case studies performing risk assessment for specific substances or groups of substances. Consortia such as EU-ToxRisk, NanoCommons and the new projects of the SC1-BHC-11-2020 programme are ideal for such cross seeding. Indeed there has been expressed interest in some of OpenRiskNet modules in EU-ToxRisk especially with respect to the read-across tool mentioned above (see “Relation to risk assessment frameworks” section).

# GLOSSARY

The list of terms or abbreviations with the definitions, used in the context of OpenRiskNet project and the e-infrastructure development is available:

https://github.com/OpenRiskNet/home/wiki/Glossary

# REFERENCES


1. Hartung T, FitzGerald RE, Jennings P, Mirams GR, Peitsch MC, Rostami-Hodjegan A, et al. Systems Toxicology: Real World Applications and Opportunities. Chem Res Toxicol. 2017;30: 870–882.
2. Jennings P, Limonciel A, Felice L, Leonard MO. An overview of transcriptional regulation in response to toxicological insult. Arch Toxicol. 2013;87: 49–72.
3. Jennings P. Stress response pathways, toxicity pathways and adverse outcome pathways. Archives of toxicology. 2013. pp. 13–14.
4. Leist M, Ghallab A, Graepel R, Marchan R, Hassan R, Bennekou SH, et al. Adverse outcome pathways: opportunities, limitations and open questions. Arch Toxicol. 2017;91: 3477–3505.
5. Pawar G, Madden JC, Ebbrell D, Firman JW, Cronin MTD. Toxicology Data Resources to Support Read-Across and (Q)SAR. Front Pharmacol. 2019;10: 561.
6. Berggren E, White A, Ouedraogo G, Paini A, Richarz A-N, Bois FY, et al. Ab initio chemical safety assessment: A workflow based on exposure considerations and non-animal methods. Comput Toxicol. 2017;4: 31–44.
7. Shuttleworth M. Case Study Research Design - How to conduct a Case Study. Available: https://explorable.com/case-study-research-design
8. ECHA. Read-Across Assessment Framework (RAAF). 2017. Available: https://echa.europa.eu/documents/10162/13628/raaf_en.pdf
9. Jennings P, Exner T, Farcal L, Oki N, Sarimveis H, Doganis P, et al. Final definition of case studies (Deliverable 1.3). 2018. doi:10.5281/zenodo.1479127
10. Gocht T, Berggren E, Ahr HJ, Cotgreave I, Cronin MTD, Daston G, et al. The SEURAT-1 approach towards animal free human safety assessment. ALTEX. 2015;32: 9–24.
11. Moretto A, Bachman A, Boobis A, Solomon KR, Pastoor TP, Wilks MF, et al. A framework for cumulative risk assessment in the 21st century. Crit Rev Toxicol. 2017;47: 85–97.
12. Blaauboer BJ, Boekelheide K, Clewell HJ, Daneshian M, Dingemans MML, Goldberg AM, et al. The use of biomarkers of toxicity for integrating in vitro hazard estimates into risk assessment for humans. ALTEX. 2012;29: 411–425.
13. Thomas RS, Philbert MA, Auerbach SS, Wetmore BA, Devito MJ, Cote I, et al. Incorporating new technologies into toxicity testing and risk assessment: moving from 21st century vision to a data-driven framework. Toxicol Sci. 2013;136: 4–18.
14. Ankley GT, Bennett RS, Erickson RJ, Hoff DJ, Hornung MW, Johnson RD, et al. Adverse outcome pathways: a conceptual framework to support ecotoxicology research and risk assessment. Environ Toxicol Chem. 2010;29: 730–741.
15. Knapen D, Vergauwen L, Villeneuve DL, Ankley GT. The potential of AOP networks for reproductive and developmental toxicity assay development. Reprod Toxicol. 2015;56: 52–55.
16. Burgdorf T, Dunst S, Ertych N, Fetz V, Violet N, Vogl S, et al. The AOP Concept: How Novel Technologies Can Support Development of Adverse Outcome Pathways. Applied In Vitro Toxicology. 2017;3: 271–277.
17. Magkoufopoulou C, Claessen SMH, Tsamou M, Jennen DGJ, Kleinjans JCS, van Delft JHM. A transcriptomics-based in vitro assay for predicting chemical genotoxicity in vivo. Carcinogenesis. 2012;33: 1421–1429.

# ANNEXES

OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

# Case Study

## Data curation and creation of pre-reasoned datasets and searching [**DataCure**]

# SUMMARY

DataCure establishes a process for data curation and annotation that makes use of APIs (eliminating the need for manual file sharing) and semantic annotations for a more systematic and reproducible data curation workflow. In this case study, users are provided with capabilities to allow access to different OpenRiskNet data sources and target specific entries in an automated fashion for the purpose of identifying data and metadata associated with a chemical in general to identify possible areas of concern or for a specific endpoint of interest (Figure 1B). The datasets can be curated using OpenRiskNet workflows developed for this case study and, in this way, cleansed e.g. for their use in model development (Figure 1A). Text mining facilities and workflows are also included for the purposes of data searching, extraction and annotation (Figure 1C).
A first step in this process was to define APIs and provide the semantic annotation for selected databases (e.g. FDA datasets, ToxCast/Tox21 and ChEMBL). During the preparation for these use cases, it became clear that the existing ontologies do not cover all requirements of the semantic interoperability layer. Nevertheless, the design of the annotation process as an online or an offline/preprocessing step forms an ancillary part of this case study even though the ontology development and improvement cannot be fully covered by OpenRiskNet and is instead organized as a collaborative activity of the complete chemical and nano risk assessment community.

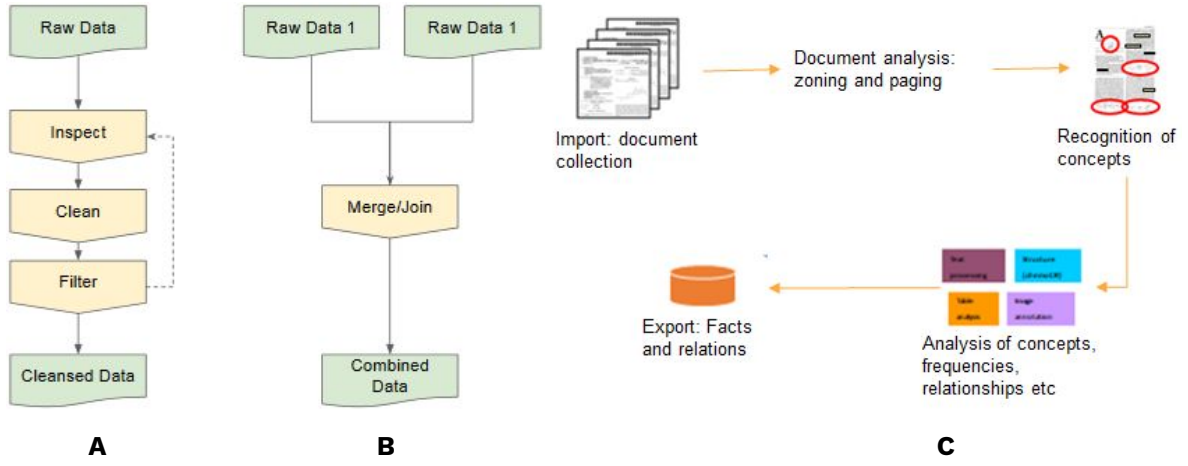



**Figure 1**. A: Raw data curation workflows; B: Data merging workflow C: Text mining workflow in general

1. **What we want to achieve?**

   Establish a process for data curation and annotation that makes use of APIs and semantic annotations for a more systematic and reproducible data curation workflow. The development of semantic annotations and API definition for selected databases are also desired.

2. **What we want to deliver?**

   The aim is to demonstrate the access to OpenRiskNet data sources, deliver curated

and annotated datasets for OpenRiskNet service users as well as preparation of and development of tools that can allow users to perform their own data curation.

3. **What is the solution and how we implement?**

   Developing resources that can make use of APIs as much as possible and eliminate the need for manual file sharing. In addition, workflows that provide examples of useful data for predictive model generation, mechanistic investigations and toxicogenomic data analysis are developed.

Expected deliverables:

- Datasets accessible via APIs
- Data extraction and curation workflows
- Data annotation

# DESCRIPTION

## Implementation team

| CS leader | Team |
|---|---|
| Noffisat Oki (EwC) | Thomas Exner (EwC), Tim Dudgeon (IM), Danyel Jennen (UM), Marc Jacobs (Fraunhofer), Marvin Martens (UM), Philip Doganis, Pantelis Karatzas (NTUA) |

## Case Study objective

- This case study serves as the entry point of collecting and curating of all data sources to be used by the remaining use cases;
- The aim is to deliver curated and annotated datasets for OpenRiskNet service users as well as preparation and development of tools and workflows that allow users to perform their own data curation and analysis;
- Semantic annotation and API definition for the selected databases are also carried out in this use case.

DataCure covers the identification of chemical/substance of concern and collection of associated existing (meta)data. The steps in Tier 0 of the underlying risk assessment framework [1] guide the data retrieval whenever applicable:

- Identification of molecular structure (if required);
- Collection of support data;
- Identification of analogues / suitability assessment and existing data;
- Identification/development of mapping identifiers;
- Collection of data for toxicology risk assessment.

# DEVELOPMENT

Steps to achieve different objectives of the DataCure, include:

- The user identifies and visualises the molecular structure:
    1. Generation of molecular identifiers for database search
    2. Searching all databases
    3. Data curation
    4. Tabular representation
    5. Visualisation
- The user collects supporting data:
    1. Provide metadata including the semantically annotated dataschema using the interoperability layer providing information on the available data
    2. Access selected databases or flat files in a directory
    3. Query to ontology metadata service and select ontologies, which should be used for annotation
    4. Annotate and index all datasets using text mining extraction infrastructure
    5. Passing to ontology reasoning infrastructure
    6. Generate database of pre-reasoned dataset (semantic integration)
    7. Allow for manual curation
- The user identifies chemical analogues:
    1. Inventory of molecules (commercially available or listed in databases)
    2. Generate list of chemically similar compounds
    3. Collect data of similar compounds

## Databases and tools

The following set of data and tools are proposed to be used and exploited within the DataCure:

- Chemical, physchem, toxicological and omics databases: PubChem, registries (e.g. ECHA, INCI), EdelweissData Explorer (EwC), Liver Toxicology Knowledge Base (LTKB) and ChEMBL
- Cheminformatics tools to e.g. convert chemical identifiers, substructure and similarity searches: RDKit, CDK, Chemical Identifier Resolver (NIH), ChemidConvert and Fragnet Search
- Ontology/terminology/annotation: SCAIView API for semantic document retrieval, JProMiner/BELIEF UIMA components for concept tagging and normalisation, TeMOwl API for identifier/name to concept mapping (Fraunhofer)
- Literature databases: PubMed, PubMed Central and ToxPlanet

## Service integration

A set of physical-chemical properties, prediction, workflows, and ontology services are integrated including the SCAIView, TeMOwl, Jaqpot, Conformal prediction, and Jupyter notebooks.

# Technical implementation

There are several steps involved in the technical implementation of this case study as listed below. All workflows mentioned here are made available to the public through the use of workflows written in scripting languages (such as R or Python) and prepared in Jupyter or Squonk notebooks. These workflows and notebooks are stored in the publicly accessible OpenRiskNet GitHub repository.

1. Data sources: EdelweissData (further described below) serves as one of the main data provisioning tools for this case study and others in the project. There are also other data sources that are directly used such as ChEMBL, PubMed and ToxPlanet, where the latter two are repositories primarily storing literature and literature based data.
2. Data Extraction: The retrieval of data from EdelweissData and all other resources are done through the use of API calls. Workflows with examples of various forms of data extraction using these APIs are documented and run from Jupyter and Squonk notebooks. Datasets can be retrieved for the query chemical and/or for similar compounds identified using the Fragnet search REST API. Additionally, these extractions partly also involve text mining using the SCAIView tool also integrated into the Jupyter notebooks.
3. Data Searching: Workflows that employ text mining capabilities are used for searching for specific data and refinement and curation of the data extracted from these sources.
4. Data curation and reasoning: This is done through the provision of workflows stored in Jupyter or Squonk notebooks. These workflows cover components such as extraction of specific data and merging of datasets for downstream analysis purposes.
5. Resubmission to data source: Even if not implemented during the OpenRiskNet project, curated datasets may be resubmitted to an OpenRiskNet-compliant data management solution like EdelweissData to be used by others through API access.

# Description of tools mentioned in technical implementation

1. EdelweissData: This platform is a web based data management tool, which is used to provide multiple OpenRiskNet data sources like ToxCast/Tox21, TG-GATEs and the Daphnia dataset. It gives users the ability to filter, search and access data based on rich metadata through an API.
2. Jupyter Notebooks: This is an open-source web application that provides an interactive programming environment to create, share, and collaborate on code and other documents.
3. Squonk Computational Notebooks: This is an open-source web application provided by IM that is somewhat similar in concept to Jupyter, but targeted at the scientist rather than the programmer.
4. SCAIView: This is an text-mining and information retrieval tool that uses semantic and ontological searches to extract relevant information from a variety of unstructured textual data sources. It can work online or offline to access both publicly available or proprietary data sources available in various formats. In the academic version three large datasets are pre-annotated: Medline (~29 mio

documents), PMC (2.6 mio) and US patent corpus (4.4 mio)

5. TeMOwl: This service provides unified access to semantic data i.e. controlled vocabularies, terminologies, ontologies and knowledge resources. It hides complexity of different semantic service providers including their various data formats. Further it aggregates (integrates, maps or aligns) different information resources. Concepts, things of thought, are often defined within multiple resources, even though they refer to the same thing. The system tries to unify those. TeMOwl is an API Service, that offers programmatic access to semantic information.
6. Fragnet Search: This is a chemical similarity search REST API that uses the Fragment Network (conceived by Astex Pharmaceuticals in this paper) to identify related molecules. The typical use of the API is to specify a query molecule and get back a set of molecules that are related to it according to some search parameters, and so is a good approach for “expanding” out one or more molecules to get other related molecules to consider. The Fragment Network is more effective and chemically meaningful compared to traditional fingerprint based similarity search techniques especially for small molecules. More information can be found here.

# OUTCOMES

Outcome from this case study provide example workflows to illustrate the processes described above and the extracted datasets with accompanying metadata were used in the AOPLink, TGX and ModelRX case studies and testing of the services integrated therein. They can now be easily adapted to real problem scenario supporting data collection in chemical risk assessment.

## Data access and curation workflow using Squonk

This Squonk computational notebook (see Figure 2) illustrates how to create a dataset of cytochrome P450 inhibition data that is collected from a number of ChEMBL datasets. Each P450 dataset (an assay in the ChEMBL database) is fetched using the ChEMBL REST API to create 4 separate datasets. These are then merged into a single dataset using the ChEMBL ID of the compound to identify molecules in common between the different datasets. The result is a tabular dataset containing the structure and the activities of the 4 different cytochrome P450s. Further processing and filtering can then be performed as needed.

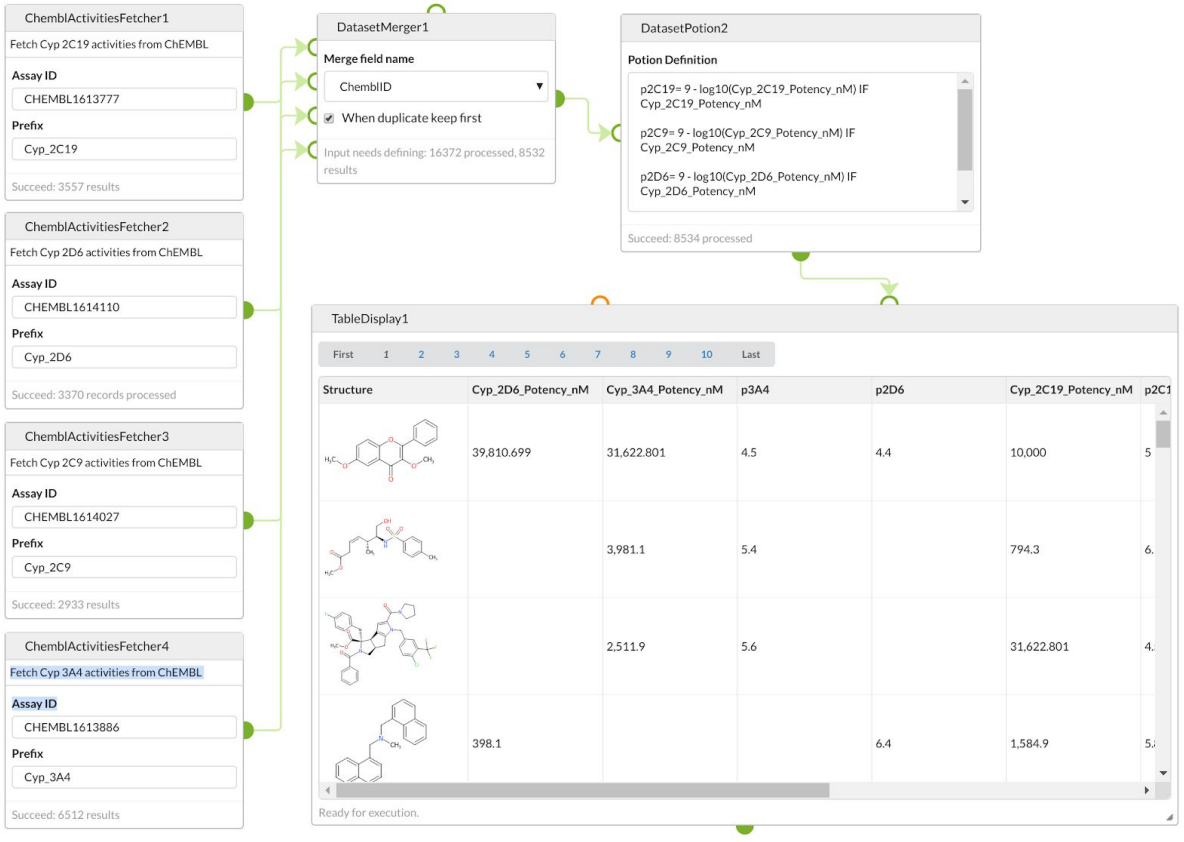



**Figure 1**. Squonk notebook to curate the P450 datasets available at https://github.com/OpenRiskNet/notebooks/tree/master/DataCure/P450

Similar notebooks could be created to fetch and merge other datasets from ChEMBL. The expectation is that these would be used for downstream activities such as

building predictive models.

## Data curation for the LTKB dataset

This Jupyter notebook illustrates approaches to preparing a dataset for use in generating predictive models. It uses the LTKB dataset from the FDA

(https://www.fda.gov/science-research/bioinformatics-tools/liver-toxicity-knowledge-base-ltkb).

Extensive data cleaning is needed as the original data is quite 'dirty'. A key aspect of the notebook is to create a 'drug-like' subset of the dataset that was used in studies performed in the ModelRX case study.

https://github.com/OpenRiskNet/notebooks/blob/master/DataCure/LTKB/LTKB_dataprep.ipynb

## Merging LTKB data with TG-GATEs

This Jupyter notebook illustrates how to join data between multiple datasets. In this case for a set of structures from TG-GATEs information from the LTKB dataset is added in order to use the DILI outcomes as a machine learning label and to use the TG-GATEs data for model generation.

https://github.com/OpenRiskNet/notebooks/blob/master/DataCure/LTKB/LTKB-TG-Gates-merge.ipynb

## Combining of data and knowledge sources

To show the benefits of making data and metadata available via semantic annotated APIs, we generated workflows to access individual OpenRiskNet data sources and more complex workflows combining and merging data from different data and knowledge sources. These were then used also as input in the AOPlink case study. Different data sources now also offer APIs to access at least part of their data (PubChem, eNanoMapper, etc.). However, these are not yet following the OpenRiskNet specifications and the API descriptions are not semantically annotated. Therefore, we will concentrate here on the resources provided by OpenRiskNet partners and especially the ToxCast/Tox21 and TG-GATEs data and the knowledge available from AOP-DB.

The EdelweissData system used to host the OpenRiskNet versions of ToxCast/Tox21 and TG-GATEs offers the full access to data and metadata through REST APIs. One example can be seen in Figure 3 using the swagger interface.

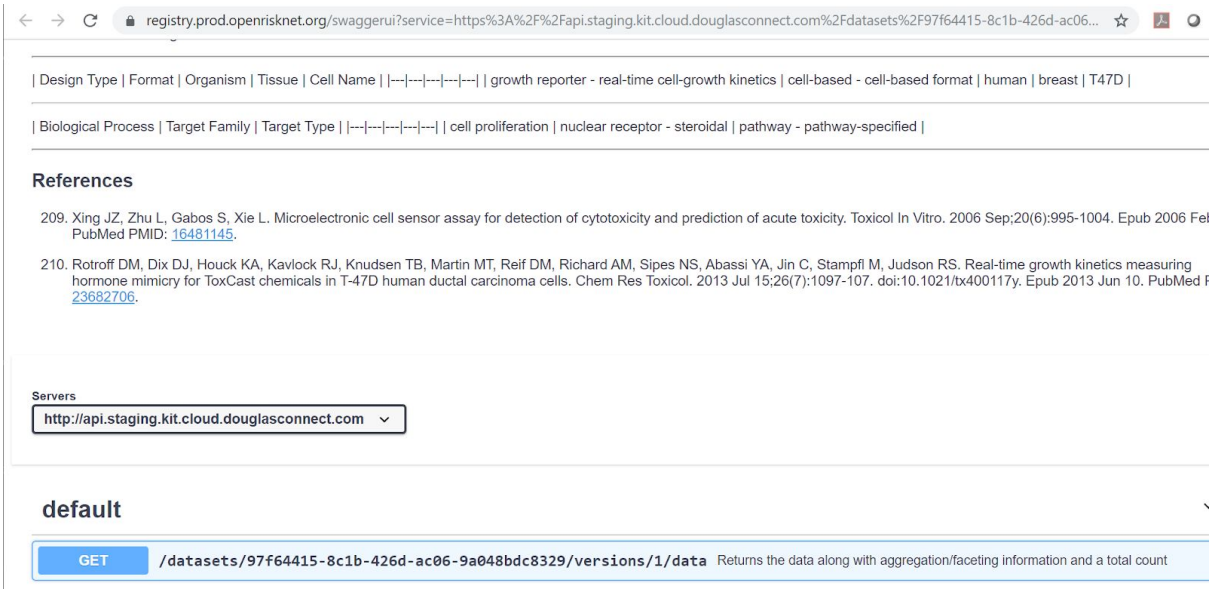



**Figure 3**. Data API for a ToxCast dataset

These APIs can be accessed using a multitude of programming languages and workflow tools. However, to simplify access, a Python library (edelweiss_data) was also developed. Access of basic metadata for all datasets for searching/browsing and of a specific dataset using this library is demonstrated in two Jupyter notebooks for ToxCast/Tox21 and TG-GATEs, respectively:

https://github.com/OpenRiskNet/notebooks/blob/master/DataCure/FetchData/AccessToxCastData.ipynb

https://github.com/OpenRiskNet/notebooks/blob/master/DataCure/FetchData/AccessTG-GatesData.ipynb

Parts of the ToxCast data access workflow are shown in Figure 4 and 5 including obtained information on all datasets and data for the first set.

## Select EdelweissData server and authenticate

[7]:
```
try:
    from edelweiss_data import API, QueryExpression as Q
except ImportError:
    pip(['install', 'edelweiss_data'])
    from edelweiss_data import API, QueryExpression as Q

edelweiss_api_url = 'https://api.staging.kit.cloud.douglasconnect.com'
api = API(edelweiss_api_url)
api.authenticate()
```

## List metadata of all ToxCast sets on the server

[8]:
```
columns = [
#    ("Endpoint", "$.assay.component.endpoint"),
    ("Endpoint name", "$.assay.component.endpoint.assay_component_endpoint_name.value"),
    ("Biological target", "$.assay.component.endpoint.target.biological_process_target.value"),
    ("Entrez gene ID for the molecular target", "$.assay.component.endpoint.target.intended.intended_target_gene.intended_target_entrez
    ("Symbol", "$.assay.component.endpoint.target.intended.intended_target_gene.intended_target_official_symbol.value"),
    ("Gene name", "$.assay.component.endpoint.target.intended.intended_target_gene.intended_target_gene_name.value")]
condition = Q.search_anywhere("EPA-ToxCast")
ToxCast = api.get_published_datasets(limit=200, columns=columns, condition=condition)
ToxCast
```

[8]:

| id | version | dataset | Endpoint name | Biological target | Entrez gene ID for the molecular target | Symbol | Gene name |
|---|---|---|---|---|---|---|---|
| 7ab126dd-3a66-4cec-938e-9121d1dd270a | 1 | <PublishedDataset '7ab126dd-3a66-4cec-938e-9121d1dd270a':1 - EPA-ToxCastV3.1-ATG_PPARa_TRANS_dn summary data> | ATG_PPARa_TRANS_dn | regulation of transcription factor activity | 5465 | PPARA | peroxisome proliferator-activated receptor alpha |
| 093c8220-a5cd-4fe0-8793-4ed032f09420 | 1 | <PublishedDataset '093c8220-a5cd-4fe0-8793-4ed032f09420':1 - EPA-ToxCastV3.1-ATG_PPARa_TRANS_up summary data> | ATG_PPARa_TRANS_up | regulation of transcription factor activity | 5465 | PPARA | peroxisome proliferator-activated receptor alpha |

**Figure 4**. Workflow to list all available ToxCast datasets. Additionally metadata can also be listed like the information on the biological target shown here.

## Access specific dataset

```
[9]: data = ToxCast.iloc[0]['dataset'].get_data()
     data
```

[9]:

| | DTXSID | DTXCID | Substance name | Substance type | Substance note | Quality control level | |
|---|---|---|---|---|---|---|---|
| 1 | None | None | 4-Hydroxynonenal | None | None | None | |
| 2 | DTXSID6057908 | None | MED_ChemMix_7EnvC | None | None | None | |
| 3 | None | None | MEDWater004_1 | None | None | None | |
| 4 | DTXSID1021455 | DTXCID401455 | FD&C Yellow 5 | Single Compound | None | database.QcLevel@2ca3970a | [Na+].[Na+].[Na+].[O-]C(=O)C1=(=O)=O)C(=O)N(N1 |
| 5 | DTXSID6020692 | DTXCID00692 | Methenamine | Single Compound | None | database.QcLevel@2ca3970a | |
| 6 | DTXSID5020154 | DTXCID30154 | Clorophene | Single Compound | None | database.QcLevel@2ca39709 | OC1=( |
| 7 | DTXSID0020654 | DTXCID80654 | Geranyl acetate | Single Compound | None | database.QcLevel@2ca39709 | |
| 8 | DTXSID4021137 | DTXCID001137 | 1,3-Benzenediamine | Single Compound | None | database.QcLevel@2ca3970a | |
| 9 | DTXSID0042400 | DTXCID8022400 | Sodium hexyldecyl sulfate | Single Compound | None | database.QcLevel@2ca3970a | [Na+].CCCC |
| 10 | DTXSID9026926 | DTXCID406926 | 1-Tetradecanol | Single Compound | None | database.QcLevel@2ca39709 | |
| 11 | DTXSID0020606 | DTXCID40606 | Bis(2-ethylhexyl)hexanedioate | Mixture of Stereoisomers | None | database.QcLevel@2ca39709 | CCCCC(CC)CC |
| 12 | DTXSID7021605 | DTXCID301605 | Hexanedioic acid | Single Compound | None | database.QcLevel@2ca39709 | |
| 13 | DTXSID7032004 | DTXCID9011121 | Flutamide | Single Compound | None | database.QcLevel@2ca3970a | CC(C)C(=O)NC1=CC |
| 14 | None | None | MEDWater004_10 | None | None | None | |

**Figure 5**. Workflow to access the first dataset in the list retrieved by the workflow in Figure 4

Case study AOPLink, identified stressors and genes related to specific key events of AOP 37: PPARalpha-dependent liver cancer. It was then investigated if these relationships can be validated with data available from OpenRiskNet sources. DataCure provided a workflow for accessing ToxCast and TG-GATEs to extract IC50 from assays with the relevant biological targets and fold changes from transcriptomics experiments, respectively. The workflow is available at:

https://github.com/OpenRiskNet/notebooks/blob/master/DataCure/FetchData/GeneSpecificData.ipynb

The data collected is shown in Figure 6 and 7. Additionally, a heatmap representing the IC50 values extracted from ToxCast are shown in Figure 8.

| | Assay | DTXSID | Substance name | InChI key | CAS | IC50 |
|---|---|---|---|---|---|---|
| **3780** | ATG_PPARa_TRANS_dn | DTXSID5020607 | Di(2-ethylhexyl) phthalate | BJQHLKABXJIVAM-UHFFFAOYSA-N | 117-81-7 | NaN |
| **3886** | ATG_PPARa_TRANS_up | DTXSID5020607 | Di(2-ethylhexyl) phthalate | BJQHLKABXJIVAM-UHFFFAOYSA-N | 117-81-7 | NaN |
| **33** | NVS_NR_hPPARa | DTXSID5020607 | Di(2-ethylhexyl) phthalate | BJQHLKABXJIVAM-UHFFFAOYSA-N | 117-81-7 | NaN |
| **936** | ATG_PPARa_TRANS_dn | DTXSID0020652 | Gemfibrozil | HEMJJKBWTPKOJG-UHFFFAOYSA-N | 25812-30-0 | NaN |
| **964** | ATG_PPARa_TRANS_up | DTXSID0020652 | Gemfibrozil | HEMJJKBWTPKOJG-UHFFFAOYSA-N | 25812-30-0 | 1.558449 |
| **3306** | ATG_PPARa_TRANS_dn | DTXSID3029869 | Bezafibrate | IIBYAHWJQTYFKB-UHFFFAOYSA-N | 41859-67-0 | NaN |
| **3407** | ATG_PPARa_TRANS_up | DTXSID3029869 | Bezafibrate | IIBYAHWJQTYFKB-UHFFFAOYSA-N | 41859-67-0 | 0.918967 |
| **4120** | ATG_PPARa_TRANS_dn | DTXSID3020336 | Clofibrate | KNHUKKLJHYUCFP-UHFFFAOYSA-N | 637-07-0 | NaN |
| **4237** | ATG_PPARa_TRANS_up | DTXSID3020336 | Clofibrate | KNHUKKLJHYUCFP-UHFFFAOYSA-N | 637-07-0 | 1.653510 |
| **413** | NVS_NR_hPPARa | DTXSID3020336 | Clofibrate | KNHUKKLJHYUCFP-UHFFFAOYSA-N | 637-07-0 | NaN |
| **2515** | ATG_PPARa_TRANS_dn | DTXSID8020331 | Ciprofibrate | KPSRODZRAIWAKH-UHFFFAOYSA-N | 52214-84-3 | NaN |
| **2590** | ATG_PPARa_TRANS_up | DTXSID8020331 | Ciprofibrate | KPSRODZRAIWAKH-UHFFFAOYSA-N | 52214-84-3 | 0.009889 |
| **135** | ATG_PPARa_TRANS_dn | DTXSID4020290 | Pirinixic acid | SZRPDCCEHVWOJX-UHFFFAOYSA-N | 50892-23-4 | NaN |
| **137** | ATG_PPARa_TRANS_up | DTXSID4020290 | Pirinixic acid | SZRPDCCEHVWOJX-UHFFFAOYSA-N | 50892-23-4 | 0.745003 |
| **480** | NVS_NR_hPPARa | DTXSID4020290 | Pirinixic acid | SZRPDCCEHVWOJX-UHFFFAOYSA-N | 50892-23-4 | 0.943310 |
| **3922** | ATG_PPARa_TRANS_dn | DTXSID2029874 | Fenofibrate | YMTINGFKWWXKFG-UHFFFAOYSA-N | 49562-28-9 | NaN |
| **4033** | ATG_PPARa_TRANS_up | DTXSID2029874 | Fenofibrate | YMTINGFKWWXKFG-UHFFFAOYSA-N | 49562-28-9 | 0.727661 |
| **295** | NVS_NR_hPPARa | DTXSID2029874 | Fenofibrate | YMTINGFKWWXKFG-UHFFFAOYSA-N | 49562-28-9 | NaN |

**Figure 6**. IC50 values of relevant ToxCast assays for the stressors of AOP 37

| | Compound | InChI Key | CAS | Organism | Organ | Study type | Dose | Duration | Duration unit | PROBEID | SYMBOL | logFC | AveExpr | t | P.Value | adj.P.Val |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **70** | gemfibrozil | HEMJJKBWTPKOJG-UHFFFAOYSA-N | [25812-30-0] | Human | Liver | in_vitro | high | 8 | hr | 206870_at | PPARA | 0.161344 | 0.716027 | 4.055956 | 4.999891e-05 | 0.007388 |
| **71** | gemfibrozil | HEMJJKBWTPKOJG-UHFFFAOYSA-N | [25812-30-0] | Human | Liver | in_vitro | high | 8 | hr | 223437_at | PPARA | 0.145851 | 1.193995 | 3.666473 | 2.461514e-04 | 0.022135 |
| **72** | gemfibrozil | HEMJJKBWTPKOJG-UHFFFAOYSA-N | [25812-30-0] | Human | Liver | in_vitro | high | 8 | hr | 1558631_at | PPARA | 0.135322 | 0.698542 | 3.401782 | 6.699593e-04 | 0.041111 |
| **73** | gemfibrozil | HEMJJKBWTPKOJG-UHFFFAOYSA-N | [25812-30-0] | Human | Liver | in_vitro | high | 8 | hr | 244689_at | PPARA | 0.127207 | 1.008897 | 3.197788 | 1.385647e-03 | 0.066398 |
| **74** | gemfibrozil | HEMJJKBWTPKOJG-UHFFFAOYSA-N | [25812-30-0] | Human | Liver | in_vitro | high | 8 | hr | 226978_at | PPARA | 0.100995 | 1.293303 | 2.538869 | 1.112390e-02 | 0.211599 |
| **75** | gemfibrozil | HEMJJKBWTPKOJG-UHFFFAOYSA-N | [25812-30-0] | Human | Liver | in_vitro | high | 8 | hr | 210771_at | PPARA | 0.064561 | 0.332642 | 1.622977 | 1.046001e-01 | 0.568553 |
| **76** | gemfibrozil | HEMJJKBWTPKOJG-UHFFFAOYSA-N | [25812-30-0] | Human | Liver | in_vitro | high | 8 | hr | 1560981_a_at | PPARA | 0.041624 | 0.506034 | 1.046356 | 2.954015e-01 | 0.783652 |
| **77** | gemfibrozil | HEMJJKBWTPKOJG-UHFFFAOYSA-N | [25812-30-0] | Human | Liver | in_vitro | high | 8 | hr | 223438_s_at | PPARA | 0.041606 | 0.901968 | 1.045919 | 2.956030e-01 | 0.783673 |
| **95** | gemfibrozil | HEMJJKBWTPKOJG-UHFFFAOYSA-N | [25812-30-0] | Human | Liver | in_vitro | low | 8 | hr | 210771_at | PPARA | -0.066571 | 0.267076 | -1.790785 | 8.392459e-02 | 0.999962 |
| **96** | gemfibrozil | HEMJJKBWTPKOJG-UHFFFAOYSA-N | [25812-30-0] | Human | Liver | in_vitro | low | 8 | hr | 1560981_a_at | PPARA | -0.045167 | 0.462639 | -1.182652 | 2.466913e-01 | 0.999962 |
| **97** | gemfibrozil | HEMJJKBWTPKOJG-UHFFFAOYSA-N | [25812-30-0] | Human | Liver | in_vitro | low | 8 | hr | 226978_at | PPARA | -0.041070 | 1.222271 | -1.084834 | 2.870602e-01 | 0.999962 |
| **98** | gemfibrozil | HEMJJKBWTPKOJG-UHFFFAOYSA-N | [25812-30-0] | Human | Liver | in_vitro | low | 8 | hr | 237142_at | PPARA | 0.040254 | -0.005626 | 1.081272 | 2.886146e-01 | 0.999962 |
| **99** | gemfibrozil | HEMJJKBWTPKOJG-UHFFFAOYSA-N | [25812-30-0] | Human | Liver | in_vitro | low | 8 | hr | 206870_at | PPARA | -0.032947 | 0.618882 | -0.885368 | 3.833461e-01 | 0.999962 |

**Figure 7**. Fold changes extracted from TG-GATEs for the stressors of AOP 37

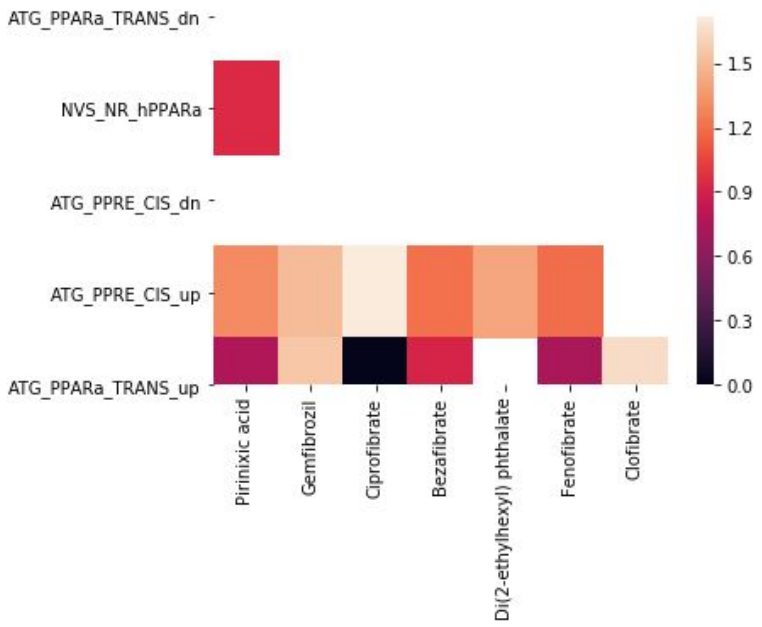



**Figure 8**. Heatmap of IC50 values extracted from ToxCast for the stressors of AOP 37

# Finding similar data-rich compounds for read across

Chemical-biological read across can be performed by searching chemically similar compounds and extracting biological data for these. Often such similarity is defined using simple chemical fingerprints. However, the FragNet approach described above can find similar compounds, which agree better with the similarity concept of a medicinal chemist. How to combine this with querying of data from ToxCast for the resulting so-called source compounds is demonstrated in the Jupyter notebook available at:

https://github.com/OpenRiskNet/notebooks/blob/master/DataCure/Fragnet/fragnet-search.ipynb

The simple molecule Piperidin-3-Amine represented by its SMILES NC1CCCNC1 is used to find similar compounds differing by one or two fragments. For these, more information is collected using the ChemidConvert OpenRiskNet services (see Figure 9).

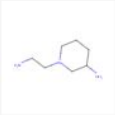
| | SMILES | Name | InChI | InChIKey |
|---|---|---|---|---|
| 0 |  | [Piperidine, 110-89-4, 571261_SIAL, 643602_ALDRICH, Piperidine on Rasta Resin, W290807_ALDRICH, AI3-24114, CCRIS 967, Cyclopentimine, Cypentil, EINECS 203-813-0, FEMA No. 2908, HSDB 114, Hexazane, Pentamethyleneimine, Pentamethylenimine, Perhydropyridine, Piperidin [German], Piperidine [UN2401] [Corrosive], Pyridine, hexahydro-, UN2401, 80645_FLUKA, Azacyclohexane, C01746, Hexahydropyridine, Piperidine, ST5213814, InChI=1/C5H11N/c1-2-4-6-5-3-1/h6H,1-5H, PIP, Piperidine solution, NCIOpen2_007828, NCIMech_000312, CHEBI:18049, 104094_SIAL, LS-3053, 411027_ALDRICH, 33537_RIEDEL, 80640_FLUKA] | InChI=1S/C5H11N/c1-2-4-6-5-3-1/h6H,1-5H2 | NQRYJNQNLNOLGT-UHFFFAOYSA-N |
| 1 |  | [1-(2-aminoethyl)piperidin-3-amine, 1-(2-aminoethyl)-3-piperidinamine, [1-(2-aminoethyl)-3-piperidyl]amine] | InChI=1S/C7H17N3/c8-3-5-10-4-1-2-7(9)6-10/h7H,1-6,8-9H2 | NQCQWIFBJYNOQM-UHFFFAOYSA-N |
| 2 |  | None | InChI=1S/C7H17N3/c8-3-5-10-4-1-2-7(9)6-10/h7H,1-6,8-9H2/t7-/m1/s1 | NQCQWIFBJYNOQM-SSDOTTSWSA-N |
| 3 |  | None | InChI=1S/C6H15N3/c7-4-6(8)2-1-3-9-5-6/h9H,1-5,7-8H2 | GQNLOVFVDPNDBK-UHFFFAOYSA-N |
| 4 |  | None | InChI=1S/C5H12N2O/c6-4-3-7-2-1-5(4)8/h4-5,7- | PSSWASGEGXCINO- |

**Figure 9**. Chemical identifiers for compounds returned by FragNet based on Piperidin-3-Amine as query

The search in the ToxCast/Tox21 datasets only results in one hit (Piperidin) even in the large Tox21 10k compound list, which shows no activity for the AhR_LUC_Agonist endpoint (see Figure 10). To focus the search on compounds for which data exists, FragNet is being optimized to additionally index the Tox21 compound list. All the neighbors found in this network will be guaranteed to be data-rich at least with respect to ToxCast/Tox21 assays.

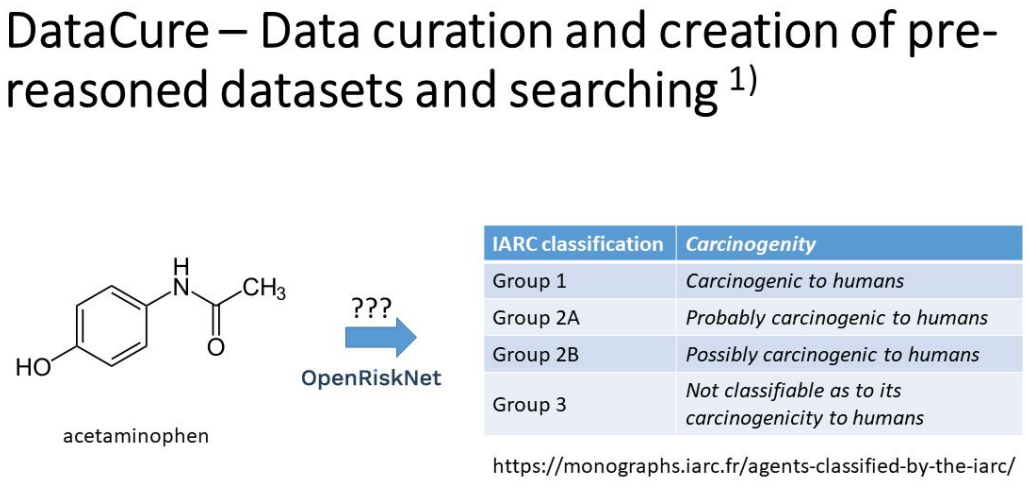
```
ToxCastData = pd.DataFrame()
for index, row in ToxCast.iterrows():
    cquery = None
    for compound in compounds['InChIKey'].values:
        if cquery is None:
            cquery = Q.fuzzy_search(Q.column('InChI key'), compound)
        else:
            cquery = cquery | Q.fuzzy_search(Q.column('InChI key'), compound)

    tmpdata = row['dataset'].get_data(condition = cquery)

    tmpdata = tmpdata[tmpdata['InChI key'].isin(compounds['InChIKey'].values)]
    tmpdata['Assay']=row['Endpoint name']
    tmpdata = tmpdata[['Assay', 'DTXSID', 'Substance name', 'InChI key', 'CAS', 'IC50']]
    ToxCastData = pd.concat([ToxCastData, tmpdata])

ToxCastData.sort_values(by=['InChI key','Assay'])
```

| | Assay | DTXSID | Substance name | InChI key | CAS | IC50 |
|---|---|---|---|---|---|---|
| 5016 | TOX21_AhR_LUC_Agonist | DTXSID6021165 | Piperidine | NQRYJNQNLNOLGT-UHFFFAOYSA-N | 110-89-4 | None |

**Figure 9**. Data for Piperidin in the AhR_LUC_Agonist assay identified as neighbor of Piperidin-3-Amine using FragNet

# Data mining workflow for carcinogenicity predictions

For this case study a text mining workflow for metadata extraction for carcinogenicity predictions was developed. This workflow outlines the implementation of a search for compound information from the literature. The main task at hand has been to find supporting information on the IARC classification from the publically available literature (i.e. journal articles) or from the ToxPlanet repository for a set of predefined compounds. Each compound should be classified as belonging to one of the four groups defined by the International Agency for Research on Cancer (IARC, see Figure 11).




1) https://openrisknet.org/e-infrastructure/development/case-studies/case-study-datacure/

**Figure 11**. Problem description: into which class belongs acetaminophen?

The IARC classification problem has been broken into several smaller tasks which can be solved by services integrated into the ORN infrastructure. All services expose an ORN compliant API, which can be accessed via an access secured API. The complete workflow has been assembled into a Jupyter Notebook. Which has also been deployed on the ORN infrastructure. This notebook outlines the implementation of a search for compound information from the literature. Workflow to be demonstrated (see also Figure 12):

1. authenticate (keycloak)
2. find proper concept to text mine ( OLS, TeMOwl )
3. find proper documents containing that concept ( SCAIView )
4. further analyze documents with NLP ( SCAIView -> UIMA ) in order to find evidence sentences supporting the classification of a compound

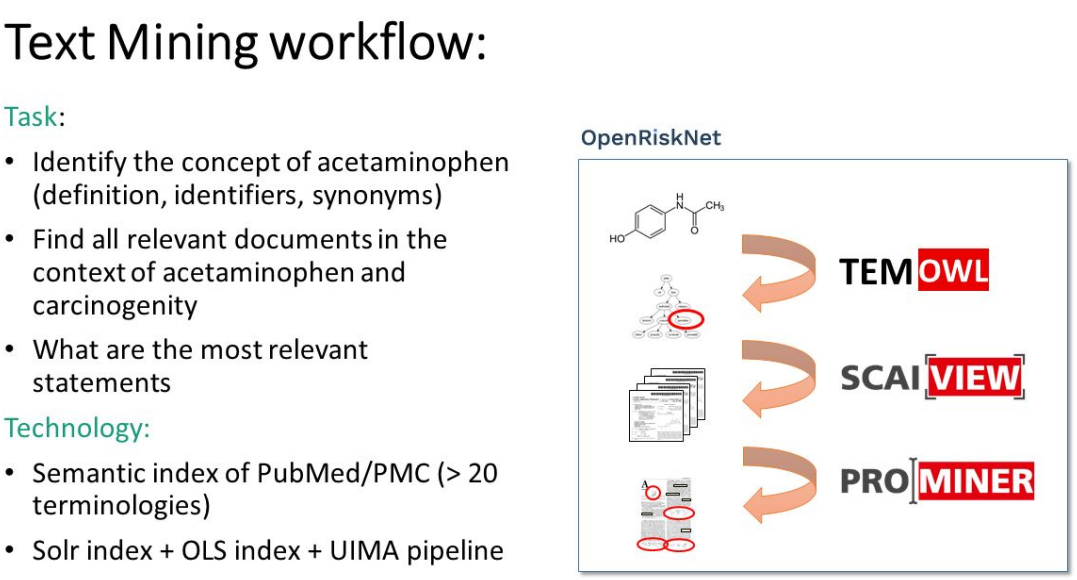



**Figure 12**. Problem description: into which class belongs acetaminophen?

In the example of acetaminophen the request to TeMOwl delivers the following information: the compound is also known as Paracetamol and more information can be retrieved via ChEBI using the identifier chebi:46195; 'A member of the class of phenols that is 4-aminophenol in which one of the hydrogens attached to the amino group has been replaced by an acetyl group.'

Searching in SCAIView for the keywords 'acetaminophen' and 'IARC' retrieves two documents from Pubmed, Searching with more general terms 'acetaminophen', 'carcinogen' and 'human' finds 8 documents. There are 51 documents talking about 'Acetaminophen', 'cancer' and 'human'. Searching in full texts instead of Pubmed abstracts finds more relevant documents eg 153 documents talking about 'acetaminophen', 'carcinogen' and 'human'.

This illustrates that we need to process the documents further to find relevant sentences since reading 153 full text documents is a time consuming and challenging task. In order to identify the wanted sentences following tools and terminologies have been used:

- DrugBank (drugs)
- Homo_sapiens (genes and proteins)
- ATC (drug classes)
- BAO (assays)
- HypothesisFinder (speculative statements)

The text mining algorithm searches for sentences which are talking about a drug or drug class in the context of humans and some cancer risk. The following list of sentences has been selected from the document with id PMC2018698: "Analgesic and anti-inflammatory drug use and risk of bladder cancer: a population based case control study"

'While some studies of bladder cancer found evidence of an elevated risk associated with heavy use of paracetamol, the majority did not, and some suggested an overall decreased

risk [[<6>,<8>,<10>,<11>,<13>-<15>,<19>], additional file<1>].',

'Our data further support an etiologic role of phenacetin in bladder cancer occurrence and they further suggest that risk increases with duration of use.Paracetamol is a metabolite of phenacetin, but it is unclear whether paracetamol retains the carcinogenic potential of its parent compound.',

'Paracetamol is not a potent inhibitor of cyclooxygenase (COX), but may inhibit NFkB, a transcription factor related to the inhibition of apoptosis [<29>], up-regulated in several cancers, including bladder cancer [<30>].',

'Metabolism of paracetamol results in a reactive metabolite (N-acetyl-P-benzoquinone imine (NAPQI)) that can form DNA adducts [<31>] and cause liver and renal toxicity [<32>].',

'Thus, paracetamol, in theory, could promote apoptosis through NFkB inhibition conferring protection against bladder cancer, or conversely, could act as a bladder carcinogen through accumulation of DNA adducts from its toxic metabolite NAPQI.',

'Recent evidence also raises the possibility of a role of genetic variation of paracetamol metabolizing genes on bladder cancer susceptibility associated with paracetamol use [<28>].',

'Further investigation of genetic variation in the metabolic pathway of paracetamol and tumor phenotype in this and other populations may help to clarify the anti-carcinogenic or carcinogenic potential of paracetamol.Aspirin and other NSAIDs are COX inhibitors (with varying isoenzyme affinities) and probably have alternative targets of action (i.e., NF kappa B inhibition) that could influence cancer occurrence [<33>].'

# DataCure webinar

Finally, a webinar demonstrating the technical implementation steps described above was given on the 18th of March 2018[1]. In this demonstration, the case study participants introduced attendees to the OpenRiskNet data handling and curation process. This included methods for data access, upload, and extraction for further downstream analysis as described in the technical implementation steps.

Specific examples demonstrated during the webinar include:

- A workflow for transcriptomics data extraction and metadata annotation from data stored in EdelweissData;
- A text mining workflow for metadata extraction for carcinogenicity predictions;
- Demonstration of extraction and curation of data for liver toxicity modeling using data from the US FDA Liver toxicity knowledgebase (LTKB).

The Jupyter notebook workflows prepared for the webinar demonstrations are available here (https://github.com/OpenRiskNet/notebooks/tree/master/DataCure/LTKB) in the OpenRiskNet GitHub.

[1] https://openrisknet.org/events/58/

# REFERENCES

1. Berggren E, et al. Ab initio chemical safety assessment: A workflow based on exposure considerations and non-animal methods. Computational Toxicology. Elsevier; 2017;4: 31–44.

OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

# Case Study

## Modelling for Prediction or Read Across [**ModelRX**]

# SUMMARY

The ModelRX case study was designed to cover the important area of generating and applying predictive models, and more specifically QSAR models in hazard assessment endorsed by different regulations, as completely *in silico* alternatives to animal testing and useful also in early research when no data is available for a compound. The QSAR development process schematically presented in Figure 1 begins by obtaining a training data set from an OpenRiskNet data source. A model can then be trained with OpenRiskNet modelling tools and the resulting models are packaged into a container, documented and ontologically annotated. To assure the quality of the models, they are validated using OECD guidelines (Jennings et al. 2018). Prediction for new compounds can be obtained using a specific model or a consensus of predictions of all models. This case study will present this workflow with the example of blood-brain-barrier (BBB) penetration, for which multiple models were generated using tools from OpenRiskNet consortium and associated partners used individually as well as in a consensus approach using Dempster-Shafer theory (Park et al. 2014; Rathman et al. 2018).

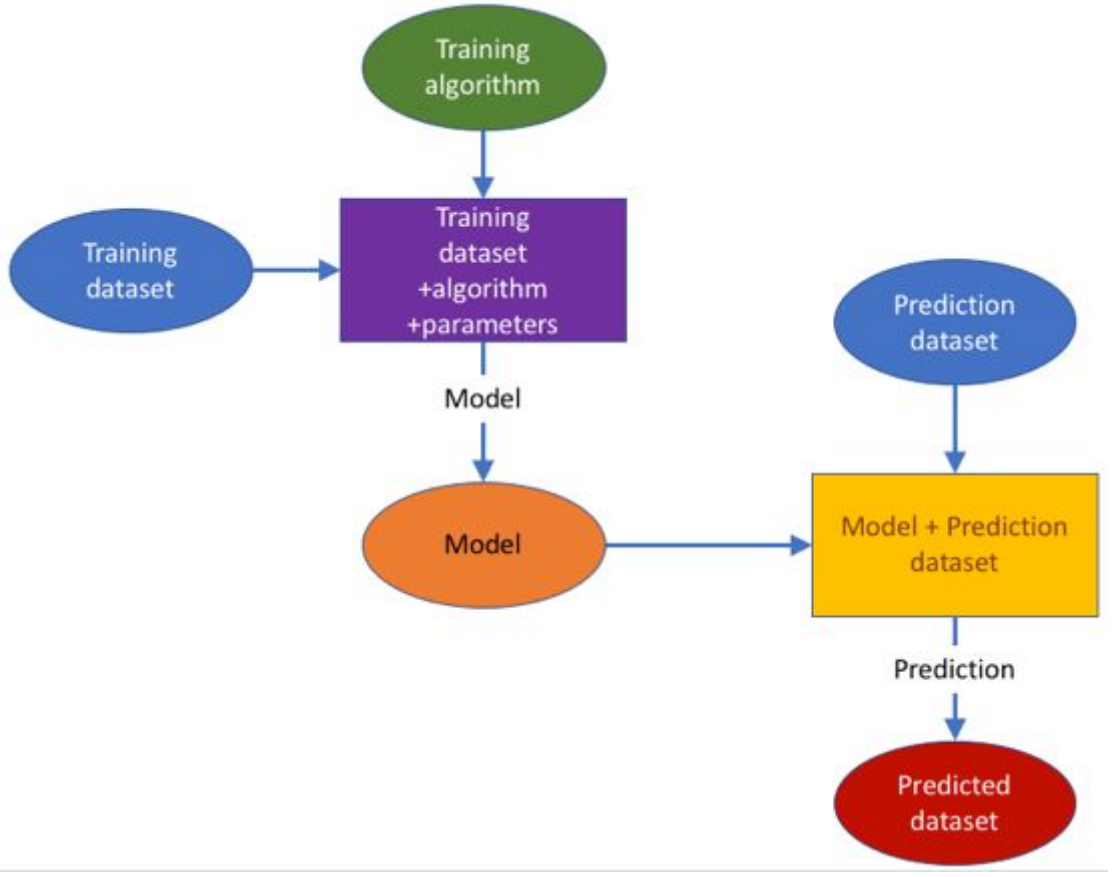



**Figure 1**. Building and using a prediction model workflow

# DESCRIPTION

## Implementation team

| CS leader | Team |
|---|---|
| Harry Sarimveis (NTUA) | NTUA, JGU, UU, EwC |

## Case Study objective

The objectives of this case study are: use *in-silico* predictive modelling approaches (QSAR) to support final risk assessment by supporting similarity identification related to the DataCure case study (by providing tools for calculating theoretical descriptors of substances) and fill data gaps for specific compounds or to augment incomplete datasets.

## Risk assessment framework

The ModelRX case study contributes in two tiers (as tiers are defined in Berggren et al. 2017):

- On the one hand, it provides computational methods to support suitability assessment of existing data and identification of analogues (Tier 0);
- On the other hand, it provides predictive modelling functionalities, which are essential in the field of final risk assessment (Tier 2).

# DEVELOPMENT

## Databases and tools

Jaqpot Quattro (NTUA), CPSign (UU), JGU WEKA Rest service (JGU), Lazar (JGU/IST).

## Technical implementation

From the developer's perspective, this case study demonstrates the improved interoperability and compatibility/complementarity of the models based on services deployed following the general steps that have been agreed for developing the OpenRiskNet infrastructure and services. Each application is delivered in the form of a container image and deployed. Docker is used as the basic engine for the containerisation of the applications. Above that, OpenShift, which is a container orchestration tool, is used for the management of the containers and services. OpenShift provides many different options for deploying applications. Some recipes and examples have been documented in the OpenRiskNet GitHub page[1].

When an application is deployed, a service discovery mechanism is responsible for discovering the most suitable services for each application. Based upon the OpenAPI specification, each API should be deployed with the swagger definition. This swagger file should then be integrated with the Json-LD annotations as dictated by the Json-LD specification. The discovery service mechanism parses the resulting Json-LD and resolves the annotations into RDF triplets. These triplets can then be queried with SPARQL. The result of the SPARQL query lets the user know which services are responsible for making models or predictions. The documentation can be found via swagger definition of each application. This way, the services are integrated into the OpenRiskNet virtual environments and can be used and incorporated into end user applications and other services as demonstrated here with a workflow performing consensus modelling.

QSAR modelling was already the main topic of the OpenTox project, which is a clear predecessor of OpenRiskNet. OpenTox had a much more focused aim and the clear goal to have very interlinked services, where it is even possible to combine parts of the workflow from different partners, e.g. descriptor calculation is performed by a service from one partner, the model is trained using algorithms from another partner, and finally the prediction is performed by integrating the trained model into a user interface of a third partner. To allow this, the technical implementation had to be based on rigorously defined modelling APIs the OpenTox standards[2]. Additionally, following these specifications and standards opened the full flexibility of model development and performing predictions to the user, which, on the one side, allows optimization of the workflow to a specific problem at hand but, on the other hand, also requires more experienced users. Therefore, the case study and the integration of QSAR services into OpenRiskNet in general used the OpenTox specifications as the starting point for developing a more flexible QSAR workflow, which, like the OpenTox workflow, has all components represented in Figure 1 but allows services to combine and simplify different steps and, in this way, e.g. reduces the necessity for rigorous usage of Uniform Resource

[1] https://github.com/OpenRiskNet/home/tree/master/openshift
[2] http://old.opentox.org/dev/apis/api-1.2

Identifiers (URIs) for simple substances, which can be referenced by chemical identifiers like SMILES or InChIs, and descriptors, e.g. when they are calculated on the fly by the service. This allows the easier integration of external tools as e.g. provided by the associated partners, which don't support all the features required for an OpenTox service. This approach allows us to include both more straightforward tools that can work together with OpenRiskNet with minimal setup, but at the same time can accommodate more feature-rich tools that require more feature-rich APIs to provide their full offering.

However, OpenRiskNet enforces additional requirements due to the broadening of the application area. As described in detail in deliverable reports D2.2 and D2.4, standardization of the APIs was not possible and even not desired to allow very different tools from many areas to be run on the OpenRiskNet virtual environments. Instead, harmonization and interoperability was obtained by semantic annotation of the APIs and the semantic interoperability later, which provide the user with the information needed to link to services using workflow tools. Modelling APIs need a high level of integration into the OpenRiskNet ecosystem. Integration with the DataCure CS is vital. On the semantic interoperability layer, training datasets should be compatible with an algorithm and, in turn, prediction datasets should be compatible with a prediction model. Additionally, in a best practice scenario, the generated models and datasets need to be accompanied with semantic metadata on their life cycle, thus enforcing semantic enrichment of the dynamically-created entities. Algorithms, models and predicted datasets are built as services, discoverable by the OpenRiskNet discovery service. This is a step that should occur whenever an entity (algorithm, model, predicted dataset) is created.

To enable the user to train models and use them in predictions, , the guidelines agreed on by OpenRiskNet for functionality, which should be provided by QSAR and read-across services to allow highest flexibility include the following steps. As already stated before, some of the requirements might become irrelevant for a specific service if it combines different steps to provide an easier way to make predictions, especially for less experienced users. More details on the features implemented in each service are annexed below.

1. **Selecting a training data set**

The user chooses among OpenRiskNet compliant data sets already accessible through the discovery service. Following the OpenTox specification, a **Dataset** includes, at a minimum:

- a dataset URI
- substances: substance URIs (each substance URI will be associated with a term from the ontology)
- features:
    - feature URIs (each feature URI will be associated with a term from the ontology)
    - values in numerical format
    - category (experimental/computed)
    - if computed, the URI of the model used to generate the values
    - Units

An alternative path to provide data in OpenRiskNet e.g. followed in the DataCure case study is by dedicated services offering the following elements:

- a dataset URI
- a semantically annotated data API providing information on how to access the data

and what specific data schema is used

With this information, it will be possible for alternative implementations to integrate into OpenRiskNet and interact with modelling services following the full OpenTox specifications, provided they also implement intermediate processing steps, i.e. within the environment of a Jupyter notebook, to structure the data so that it fulfills the minimum set of requirements for the following steps of the QSAR workflow.

2. **Selecting a (suitable) modelling algorithm**

The user chooses a suitable algorithm. **Algorithms** include at a minimum:

- algorithm URI
- title
- description
- algorithm type (regression/classification)
- default values for its parameters (where applicable)

3. **Specifying parameters and Generating Predictive model**

Once an algorithm has been selected, the user defines the endpoint, selects the tuning parameters, (only if different values from the default ones are desired) and runs the algorithm. The generated **Model** contains, at a minimum:

- model URI
- title
- description
- the URI of the dataset that was used to create it
- the URIs of the input features
- the URI of the predicted feature
- values of tuning parameters

The following were identified as possible extensions:

- *Include services/APIs for validation of the generated model*

  This has been implemented by Jaqpot (NTUA) as well as Weka (JGU) and Lazar (JGU).

- *Provide mechanisms to pick out the best algorithm for a specific dataset: (e.g. RRegrs)*

  As this is a highly resource-intensive process and requires significant exploration of possible choices and parameters, it would translate into additional workload on the infrastructure, which would be challenging to sustain. That, together with the fact that users have a selection of tools in the search for the best model: tools like RRegrs[3] (written in the R language) and TPOT (written in Python) can be used in common R-Python notebooks so that the user settles down on the most appropriate model and then resorts to OpenRisket tools.

[3] https://github.com/muntisa/RRegrs

- *Include algorithms to calculate domain of applicability*

  Domain of applicability calculations have been implemented by Jaqpot (NTUA). An alternative to applicability domains using Conformal Prediction methodologies (Norinder et al, 2014) is provided by the CPSign/ModelingWeb tool (UU).

**4. Selecting a prediction data set**

After the creation of a model, the user selects a prediction dataset meeting all the requirements specified in (Chomenidis et al. 2017). This dataset is tested for compatibility against the required features of the model in terms of feature URIs, i.e. the dataset should contain all the subset of features used to produce the model. Additional features are allowed, however they will be ignored.

**5. Running predictive model on the prediction data set**

The predictive model is applied on the prediction dataset to generate the predicted dataset, which must be compatible with the requirements specified in (Chomenidis et al. 2017). The predicted dataset augments the prediction dataset with all necessary information about the predicted feature:

- prediction feature URIs (each feature URI will be associated with a term from the ontology)
- values in numerical format
- category (computed)
- the URI of the model used to generate the values
- units

# OUTCOMES

The work on the case study was designed to showcase how the workflow defined above for producing semantically annotated predictive models can be shared, tested, validated and eventually be applied for predicting adverse effects of substances in a safe by design and/or risk assessment regulatory framework. OpenRiskNet provides the necessary functionalities that allow not only service developers but also researchers and practitioners to easily produce and expose their models as ready-to-use web applications. The OpenRiskNet e-infrastructure serves as a central model repository in the area of predictive toxicology. For example, when a research group publishes a predictive model in a scientific journal, they can additionally provide the implementation of the model as a web service using the OpenRiskNet implementation. The produced models contain all the necessary metadata and ontological information to make them easily searchable by the users and systematically and rigorously define their domain of applicability. Most importantly, the produced resources are not just static representations of the models, but actual web applications where the users can supply the necessary information for query substances and receive the predictions for their adverse effects. However, because of the harmonization and interoperability, these are not just stand-alone tools but can be easily combined to improve the overall performance or can be used to replace older tools with newer ones without changing the overall procedure. This was demonstrated with the workflows for developing a consensus model for blood-brain-barrier (BBB) penetration (available online[4]). The test set of 414 compounds was obtained from the Lazar service[5]. Part of this dataset is shown below:

[4] https://github.com/OpenRiskNet/notebooks/tree/master/ModelRX/Blood-brain%20barrier%20-%20Consensus

[5] https://lazar.prod.openrisknet.org/predict/dataset/blood-brain-barrier

| SMILES | Blood-Brain-Barrier Penetration |
|---|---|
| OC[C@](c1onc(n1)c1ncn2-c3cccc(c3C(=O)N(Cc12)C)Cl)(O)C | non-penetrating |
| NCCc1nc2n(c1)cccc2 | non-penetrating |
| NCCc1nc2n(c1)cccc2 | non-penetrating |
| CCCN(CCC)CCc1ccc(c2c1CC(=C)N2)O | penetrating |
| Fc1ccc2c(c1)onc2C1CCN(CC1)CCc1c(C)nc2n(c1=O)CCC[C@H]2O | penetrating |
| Clc1ccc2-n3c(CN=C(c2c1)c1ccccc1)nnc3C | penetrating |
| CN(CC/C=C/1\c2ccccc2CCc2c1cccc2)C | penetrating |
| Oc1ncnc2c1cn[nH]2 | penetrating |
| O=c1cc(n(n1c1ccccc1)C)C | penetrating |
| CC(=O)Oc1ccccc1C(=O)O | penetrating |
| ClCCNC(=O)N(CCCl)N=O | penetrating |
| ClCCNC(=O)N(CCCl)N=O | penetrating |
| Cn1cnc2c1c(=O)n(C)c(=O)n2C | penetrating |
| NC(=O)N1c2ccccc2C=Cc2c1cccc2 | penetrating |
| ClCCN(c1ccc(cc1)CCCC(=O)O)CCCl | non-penetrating |
| CN(CCCN1c2ccccc2Sc2c1cc(Cl)cc2)C | penetrating |
| N#C/[NH]=C(/NCCSCc1nc[nH]c1C)\NC | non-penetrating |
| Clc1cccc(c1NC1=NCCN1)Cl | penetrating |
| COc1ccc2c3c1O[C@@H]1[C@@]43CCN([C@H](C2)[C@@H]4C=C[C@@H]1O)C | penetrating |
| CNCCCN1c2ccccc2CCc2c1cccc2 | penetrating |
| OC[C@@H]1CC[C@@H](O1)n1cnc2c1ncnc2O | non-penetrating |
| Clc1ccc2c(c1)[nH]c(=O)n2C1CCN(CC1)CCCn1c(=O)[nH]c2c1cccc2 | penetrating |
| OCCOCCN1CCN(CC1)[C@@H](c1ccc(cc1)Cl)c1ccccc1 | penetrating |
| CC(Cc1ccc(cc1)[C@@H](C(=O)O)C)C | penetrating |
| COc1cccnc1CCCCNc1[nH]cc(c(=O)n1)Cc1ccc(nc1)C | non-penetrating |

In the consensus modelling, we combine independent sources of evidence in a well defined manner to generate the final prediction and estimate its uncertainty. For that purpose we employed the Dempster-Shafer theory (DST), which provides a solid mathematical framework for combining multiple evidences, where each of them is characterized by evidence-specific certainty (Park et al. 2014, Rathman et al. 2018).

Here we combine multiple independent *in silico* predictive models, each of which classifies compounds as BBB penetrating or non-penetrating. The models, which were available before the start of the project or were specifically generated for this case study, are:

**LAZAR (JGU/IST):** The model is based on the MP2D fingerprints and uses nearest-neighbour algorithm with a weighted majority vote and distance based on the Tanimoto similarity with a threshold of 0.1. The model was validated using 3-fold cross-validation.

**Jaqpot (NTUA):** The model is based on the Mordred descriptors obtained from the SMILES using RDKit and recursive feature elimination for the selection of the 20 most important features. The model uses logistic regression based on these 20 features. The model was validated using 10-fold cross-validation.

**CPSign (UU):** The Cross Venn-ABERS predictor (CVAP), as implemented in the CPSign software. The signatures descriptor of atom neighbours height 1 to 3 was used together with an SVM with RBF kernel as underlying learning model. The gamma and cost values were optimized with 10-fold cross-validation and set to 0.0039 (gamma) and 4 (cost).

**JGU WEKA Rest service (JGU):** The model is based on features/fingerprints

extracted from the Blood-Brain Barrier dataset using a graph mining based algorithm called LAST-PM or Latent Structure Pattern Mining. The fingerprints used in the majority of chemical toolkits are handcrafted by chemical experts, however, identifying frequent or correlated subgraphs has the potential to reveal latent information not present in any individual ground features. This also provides a potentially different perspective to the problem at hand. The model is created using Support Vector Machines (specifically the implementation provided by LibSVM in Weka). The model parameters (cost and gamma) have been tuned using grid search.

**OCHEM (BIGCHEM):** On-line Chemical Database and Modeling environment (OCHEM) platform was used to develop OCHEM model, which was based on Associative Neural Network (ASNN, Tetko 2008) and alvaDesc (https://www.alvascience.com/alvadesc/) descriptors. The default hyper-parameters of the ASNN were used. Namely, neural networks with one hidden layer, which included three neurons each, were used. Each network was trained for 1000 iterations using early-stopping. ASNN used an ensemble of 64 individual networks, predictions of which were averaged, to provide final model predictions. The 2D to 3D conversion was done using Corina (https://www.mn-am.com/) program. The developed model is available at http://ochem.eu/model/12147752. The accuracy of predictions are estimated based on the uncertainty of ensemble predictions as described in (Sushko et al 2010).

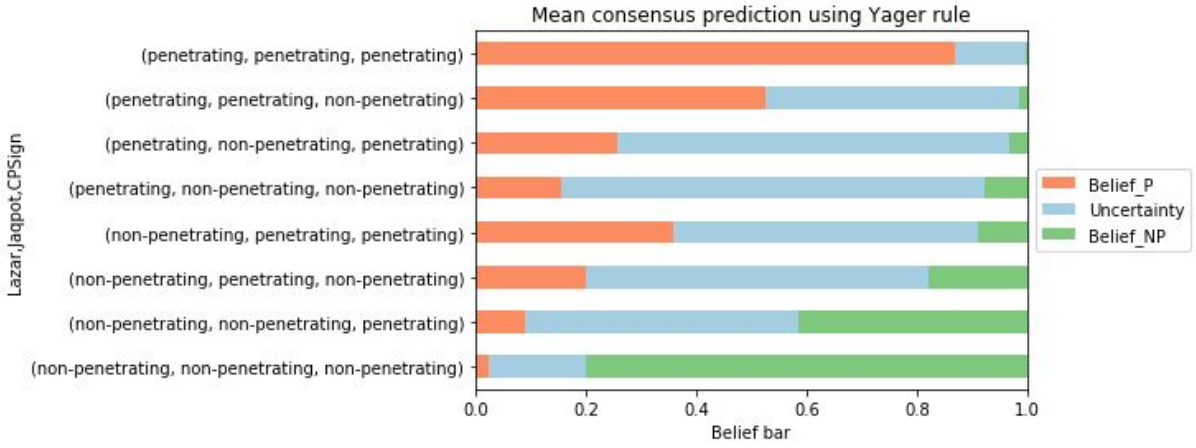
In DST, each model used for the consensus is not weighted equally but based on the confidence in the individual predictions. The certainty of a positive or negative prediction of every model is characterized by its positive or negative predictive value (PPV or NPV), where PPV is defined as the fraction of true positives of all the positive predictions (PPV = TP / (TP + FP)), and NPV is defined as the fraction of true negatives of all the negative predictions by a given model (NPV = TN / (TN + FN)). In our case PPV and NPV were obtained from the k-fold cross-validation of the respective models. Since the models from the associated partners were not completely integrated into the OpenRiskNet infrastructure at the time of this writing, only the LAZAR, Jaqpot, CPSign, WEKA and OCHEM models are considered in the following analysis.

| | LAZAR | Jaqpot | CPSign | WEKA | OCHEM |
|---|---|---|---|---|---|
| **PPV** | 0.765 | 0.864 | 0.806 | 0.909 | 0.86 |
| **NPV** | 0.690 | 0.788 | 0.698 | 0.924 | 0.71 |

DST additionally allows the user to put more or less weight on the concordance between the sources of evidence by the choice of the so-called combination rule. In this way the user is allowed to choose how conservative the consensus prediction should be. The two rules examined in this case study are Dempster and Yager combination rule. The first one neglects the disagreements among the various sources of evidence and provides lower uncertainties, whereas the conflicts among the sources of evidence result to a greater uncertainty when using the latter one. In other words, Dempster combination rule provides results similar to the majority voting rule, whereas Yager rule is more likely to

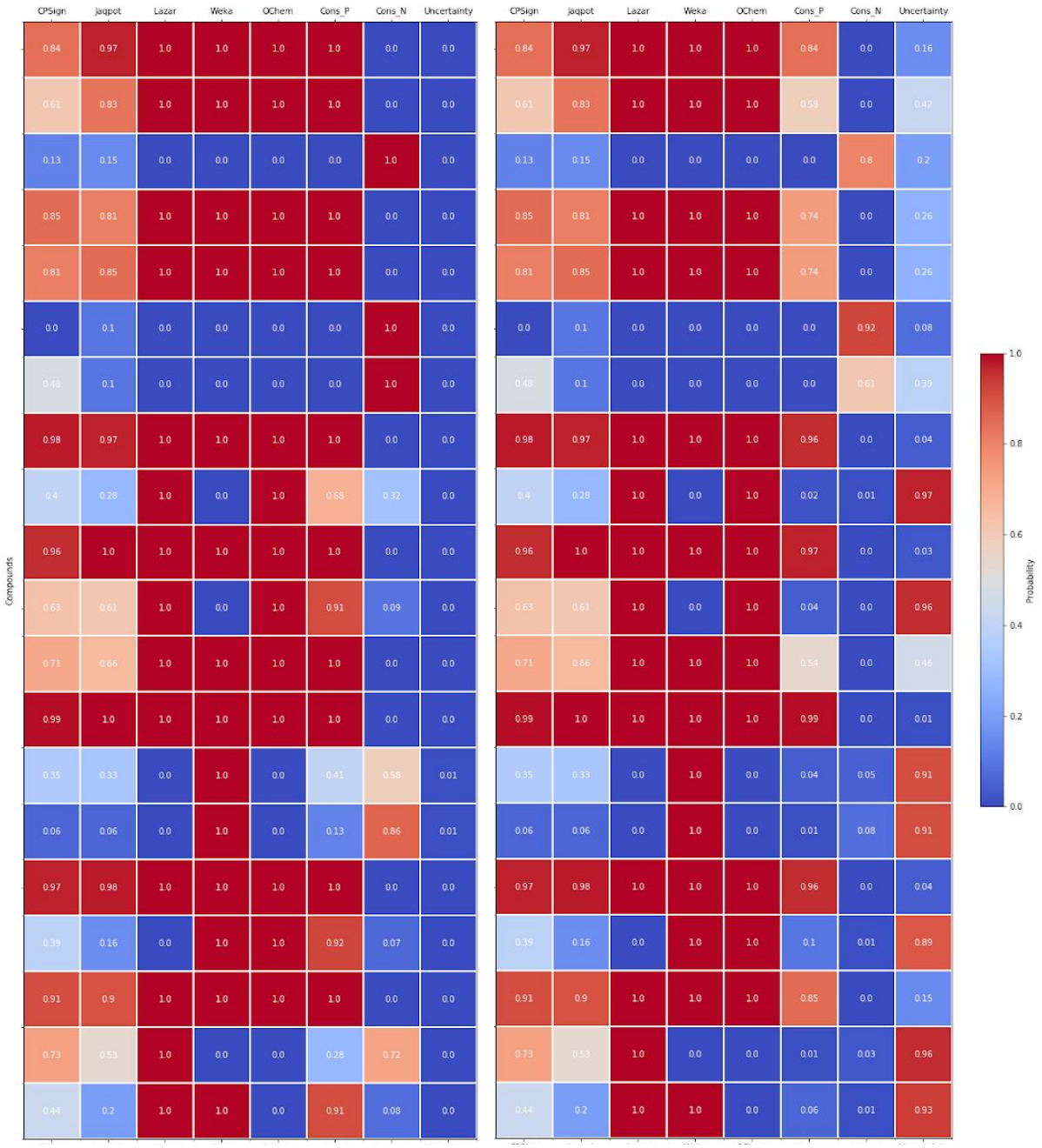
produce equivocal prediction in case of disagreements between the sources of evidence.

For every possible outcome, the DST provides **belief** and **plausibility**, which can be viewed as the lower and upper bound of the probability of that outcome, respectively, and their difference is the **uncertainty** of that outcome.
To illustrate these concepts, the figure below shows the computed average belief, plausibility and uncertainty for different combinations of predictions from three sources of evidence (where each source of evidence is a single predictive model, which predicts a compound to be penetrating or non-penetrating). The three models used for this demonstration are Lazar (JGU/IST), Jaqpot (NTUA) and CPSign (UU). The red bars represent the belief that a compound is penetrating, green bars show that it is non-penetrating and the blue bar is the corresponding uncertainty.
Belief of penetrating and uncertainty sum up to the plausibility that compound is penetrating. Analogously, the belief of non-penetrating and uncertainty sum up to plausibility that a compound is non-penetrating. Clearly, both beliefs and uncertainty should always sum up to 1.




The heatmaps below show the difference of the consensus predictions using the Dempster rule (left) and the Yager rule (right). The color designates the probability and ranges from blue (0) through white (0.5) to red (1). Every row represents the result for a single compound. The first 4 columns in each figure correspond to the probability that a given compound is BBB penetrating according to the individual predictive models. The next two columns (termed *Cons_P* and *Cons_N*) represent the consensus belief that a given compound is penetrating (*Cons_P*) or non-penetrating (*Cons_N*), while the last column depicts the uncertainty of the consensus prediction. Note that *Cons_P*, *Cons_N* and *Uncertainty* always sum up to 1.

When all 4 models agree in their prediction (all 4 are blue or red), the consensus prediction is very clear. However, the difference between the two combination rules becomes evident when we look at the cases where the models disagree. Clearly, Dempster rule tends to diminish uncertainty and puts more weight to the prevailing prediction, while the Yager rule tends to increase uncertainty, being a more conservative prediction since more molecules will show high uncertainty levels preventing a clear categorization.



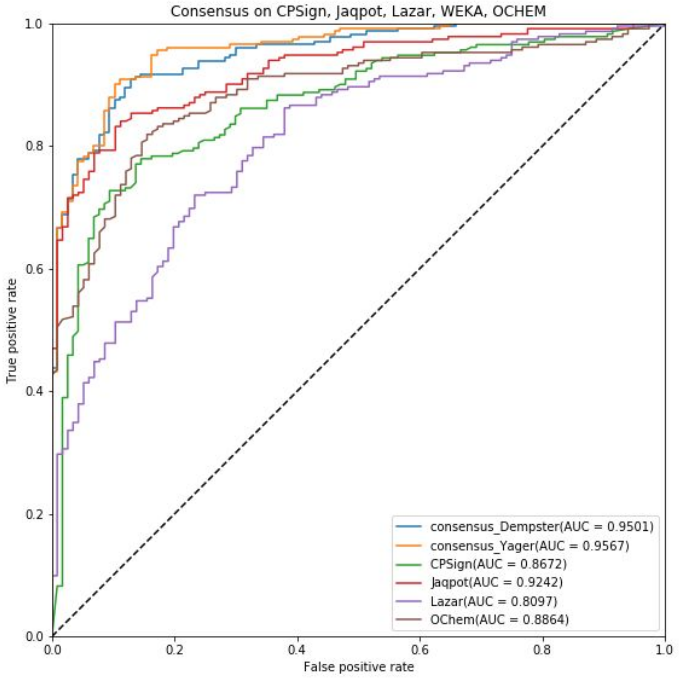

Heatmap representation of BBB predictions of individual and consensus models. Left side: consensus model using the Dempster combination rule. Right side: consensus model using the Yager combination rule. Each row presents the results for a single compound. Only the results for the first 20 compounds are displayed. Columns 1 to 5 represent the probability that a given compound is BBB penetrating as predicted by individual models (from left to right: CPSign, Jaqpot, Lazar, WEKA, OCHEM). Columns 6 and 7 represent the consensus probability that a given compound is BBB penetrating and non-penetrating, while column 8 depicts the uncertainty of the consensus approach. The figure was generated within the workflow of the *batch-compounds-offline.ipynb* Jupyter notebook available over GitHub[6].

6 https://github.com/OpenRiskNet/notebooks/blob/master/ModelRX/Blood-brain%20barrier%20-%20Consensus/batch-compounds-offline.ipynb

The performance of individual predictive models and the consensus predictions can be compared on the ROC plot (below). Clearly, the consensus predictions based on five models offer better performance, which is also reflected in greater AUC values. Note that for the WEKA predictive model the results could not be quantified, so the ROC curve could not be made.

Also it should be noted that not every predictive model was able to make a prediction for all the compounds of the test set. For example:
- The Lazar predictive model uses a nearest-neighbour algorithm with a cutoff chemical similarity of 0.1. If there was no compound found in the training set that would be sufficiently similar, then prediction could not be made.
- The Jaqpot predictive model uses Mordred descriptors, which are generated from the Mol representation of the compound. During preprocessing the Mol representations were generated from the SMILES strings using RDKit, but not all conversions were successful. Hence in the analysis for Jaqpot compounds for which Mol representation and Mordred descriptors could not be calculated were removed.

These restrictions resulted in the final number of 348 compounds, for which the consensus predictions could be made.




This figure was generated within the workflow of the *model-comparison.ipynb* Jupyter notebook available over GitHub[7]. More complete summary and comparison of consensus predictions of different combinations of models is available in the *model-comparison-summary.ipynb* also available over GitHub[8].

7 https://github.com/OpenRiskNet/notebooks/blob/master/ModelRX/Blood-brain%20barrier%20-%20Consensus/model-comparison.ipynb.

8 https://github.com/OpenRiskNet/notebooks/blob/master/ModelRX/Blood-brain%20barrier%20-%20Consensus/model-comparison-summary.ipynb

# REFERENCES


Berggren, Elisabet, Andrew White, Gladys Ouedraogo, Alicia Paini, Andrea-Nicole Richarz, Frederic Y. Bois, Thomas Exner, et al. 2017. “Ab Initio Chemical Safety Assessment: A Workflow Based on Exposure Considerations and Non-Animal Methods.” *Computational Toxicology (Amsterdam, Netherlands)* 4 (November): 31–44.

Chomenidis, Charalampos, Georgios Drakakis, Georgia Tsiliki, Evangelia Anagnostopoulou, Angelos Valsamis, Philip Doganis, Pantelis Sopasakis, and Haralambos Sarimveis. 2017. “Jaqpot Quattro: A Novel Computational Web Platform for Modeling and Analysis in Nanoinformatics.” *Journal of Chemical Information and Modeling* 57 (9): 2161–72.

Gajewicz, Agnieszka, Nicole Schaeublin, Bakhtiyor Rasulev, Saber Hussain, Danuta Leszczynska, Tomasz Puzyn, and Jerzy Leszczynski. 2015. “Towards Understanding Mechanisms Governing Cytotoxicity of Metal Oxides Nanoparticles: Hints from Nano-QSAR Studies.” *Nanotoxicology* 9 (3): 313–25.

Jennings, Paul, Thomas Exner, Lucian Farcal, Noffisat Oki, Harry Sarimveis, Philip Doganis, Danyel Jennen, et al. 2018. “Final Definition of Case Studies (Deliverable 1.3),” November. https://doi.org/10.5281/zenodo.1479127.

Norinder, Ulf, Carlsson, Lars, Boyer, Scott, Eklund, Martin. 2014. “Introducing conformal prediction in predictive modeling. A transparent and flexible alternative to applicability domain determination”. *Journal of Chemical Information and Modeling* 23;54(6):1596-603. https://doi.org/10.1021/ci5001168.

Park, Sung Jin, Ogunseitan, Oladele A, Lejano, Raul P. 2014. “Dempster-Shafer theory applied to regulatory decision process for selecting safer alternatives to toxic chemicals in consumer products”. *Integrated Environmental Assessment and Management* 10 (1): 12-21. https://doi.org/10.1002/ieam.1460

Rathman, James F, Yang, Chihae, Zhou, Haojin. 2018. “Dempster-Shafer theory for combining in silico evidence and estimating uncertainty in chemical risk assessment”. *Computational Toxicology* 6: 16-31. http://dx.doi.org/10.1016/j.comtox.2018.03.001

Tetko, Igor V. 2008. “Associative neural network”. *Methods in Molecular Biology* 458: 185-202. doi:10.1007/978-1-60327-101-1_10

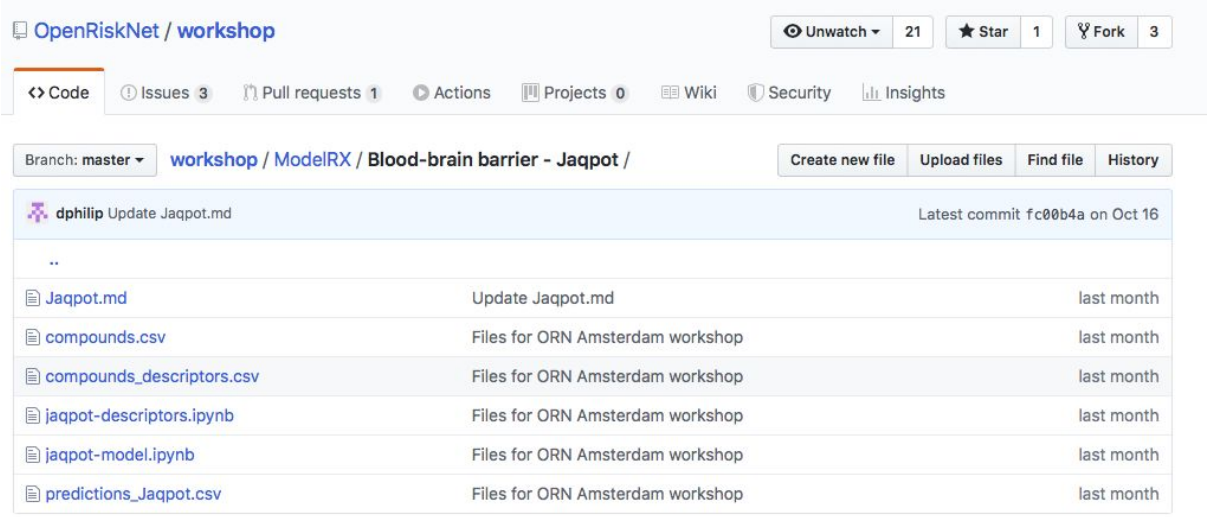
Sushko, Iurii, Novotarskyi, Sergii, Körner, Robert, Pandey, Anil Kumar, Kovalishyn, Vasily V, Prokopenko Volodymyr V, Tetko Igor V. 2010. “Applicability domain for in silico models to achieve accuracy of experimental measurements”. *Journal of Chemometrics* 24 (3-4): 202-208. https://doi.org/10.1002/cem.1296

# APPENDIX

## Jaqpot

Jaqpot (developed at NTUA) allows users to transform models they create in Python into web services with 1 line of code using the Jaqpotpy package. For more extended reference to the Jaqpotpy package, please refer to the package reference at https://jaqpotpy.readthedocs.io and for Jaqpot in PBPK modelling, please refer to the REVK Case Study.

In order to briefly demonstrate the functionality of Jaqpot, we will present an example of the user creating a new model in Python, making the model as a web service in Jaqpot and finally, using the new model to make predictions. The example was presented at the Modelling Session of the Final OpenRiskNet workshop in Amsterdam and is available here: https://github.com/OpenRiskNet/workshop/tree/master/ModelRX (please note that all services run on the cloud and the CSV files are only provided to users for comparison purposes).




The example is split in two phases, with respective notebooks. In the first phase (in the https://github.com/OpenRiskNet/workshop/blob/master/ModelRX/Blood-brain%20barrier%20-%20Jaqpot/jaqpot-descriptors.ipynb notebook), users first get the dataset from Lazar:

**Communicate with Lazar to obtain the dataset**

```
13]:
url = 'https://lazar.prod.openrisknet.org/endpoint'
headers = {'accept': 'application/json',
           'Content-Type': 'application/x-www-form-urlencoded'}

r1 = requests.get(url, headers=headers)

print("LAZAR Status code GET endpoints: {0}".format(r1.status_code))
if r1.status_code == 200:
    endpoints = r1.json()
```

```
LAZAR Status code GET endpoints: 200
```

After necessary preprocessing steps on the dataset are performed, the user produces Mordred descriptors on the dataset:

**Calculate Mordred descriptors**

```
[31]: calc = Calculator(descriptors)

dfMord = calc.pandas(df['Mol'])

dfMord.head()
```

[31]:

| | ABC | ABCGG | nAcid | nBase | SpAbs_A | SpMax_A | SpDiam_A | SpAD_A | SpMAD_A | LogEE_A | ... | SRW10 | TSRW10 | MW | AMW | WPath | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 21.474080 | 17.978542 | 0 | 0 | 34.5534 | 2.54198 | 4.93359 | 34.5534 | 1.27976 | 4.25118 | ... | 10.428837 | 78.871649 | 389.089082 | 9.048583 | 1727 | |
| 1 | 9.151948 | 8.206878 | 0 | 1 | 15.659 | 2.37835 | 4.57188 | 15.659 | 1.30491 | 3.42249 | ... | 9.190852 | 56.587917 | 161.095297 | 7.004143 | 197 | |
| 2 | 9.151948 | 8.206878 | 0 | 1 | 15.659 | 2.37835 | 4.57188 | 15.659 | 1.30491 | 3.42249 | ... | 9.190852 | 56.587917 | 161.095297 | 7.004143 | 197 | |
| 3 | 14.946702 | 13.140670 | 0 | 1 | 25.0359 | 2.45245 | 4.79766 | 25.0359 | 1.2518 | 3.90305 | ... | 9.742908 | 67.137495 | 274.204513 | 5.960968 | 862 | |
| 4 | 24.862776 | 17.808737 | 0 | 1 | 40.9336 | 2.46674 | 4.9288 | 40.9336 | 1.32044 | 4.38836 | ... | 10.513824 | 81.350168 | 426.206719 | 7.348392 | 3047 | |

5 rows × 1826 columns

The results are in turn processed to make sure only meaningful results will be used for modelling.

In the second phase, users build their model (available as a Python notebook at https://github.com/OpenRiskNet/workshop/blob/master/ModelRX/Blood-brain%20barrier%20-%20Jaqpot/jaqpot-model.ipynb).

Users develop a predictive model for blood-brain-barrier penetration using Logistic Regression (from scikit-learn[9]), after reducing the number of descriptors used to 20, based on Recursive Feature Elimination (scikit-learn).

**Split the main dataframe into X and Y**

```
[5]: X = df.drop(['True', 'SMILES'], axis=1)
Y = df[['True']].replace({'non-penetrating': 0, 'penetrating': 1})
```

**Scaling between 0 and 1**

```
[6]: X_scaled = (X - X.min()) / (X.max() - X.min())
```

**Select only 20 most important features**

More here: Recursive Feature Elimination (scikit-learn)

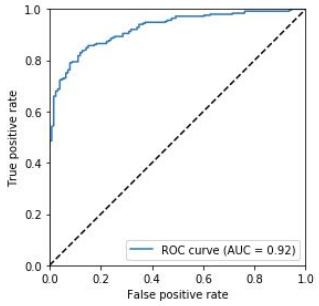
https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

```
[7]: model = LogisticRegression(solver='newton-cg', multi_class='multinomial', max_iter=100)

rfe = RFE(model, 20, verbose=1)
fit = rfe.fit(X_scaled, Y['True'])

X_rfe = X_scaled.loc[:, fit.support_.tolist()]
```

After evaluating model performance through the Confusion matrix, the Positive predictive value, the Negative predictive value and the ROC plot (shown below) the model is accepted.

[9] https://scikit-learn.org

```
ax.legend(loc="lower right")
ax.set_xlabel('False positive rate')
ax.set_ylabel('True positive rate')
ax.set_aspect('equal')
ax.set_xlim([0, 1])
ax.set_ylim([0, 1])

fig.tight_layout()

plt.show()
```




The trained model can be added to Jaqpot using the Jaqpotpy library that was specially written by the NTUA team for the purpose of uploading models to Jaqpot. First the user needs to define which Jaqpot instance actions will refer to and authenticate access to it.

### Deploy the model to Jaqpot service

```
[28]: # URL to access the jaqpot service
jaqpot = Jaqpot("https://api-jaqpot.prod.openrisknet.org/jaqpot/services/")

#alternative link for Jaqpot services
#jaqpot = Jaqpot("https://api.jaqpot.org/jaqpot/services/")
```

### Authentication ¶

You can authenticate yourself either through username/password (have to entered everytime) or with API key obtained from the Jaqpot website (requires login and has validity for a limited time - does not have to entered everytime)

```
[ ]: #with username/password
#jaqpot.request_key_safe()

#with API key
apiKey = "eyJhbGciOiJSUzI1NiIsInR5cCIgOiAiSldUIiwia2lkIiA6ICJoX2p2Z3I3bWZ4VGJ3OHJLNW9Fb3dWWUVHUms2Z0hsLW9sSjdPUnQ3V2QwIn0.eyJqdGkiOiJ
jaqpot.set_api_key(apiKey)
```

After that, uploading a ready model is a matter of 1 line of code, where the user defines the details for the model and necessary references:

### From model to web service in 1 line of code

```
[ ]: # deploy the model to jaqpot
url = jaqpot.deploy_linear_model(model, X_rfe, Y[['True']], title="OpenRiskNet/ModelRX", description="Logistic regression model + RFE
```

You now have a web service.

Now that the model has become a web service, getting predictions from it is simple:

You can also get predictions from your model:

```
[ ]: pred, predCol = jaqpot.predict(X_rfe, modelId=url)
```

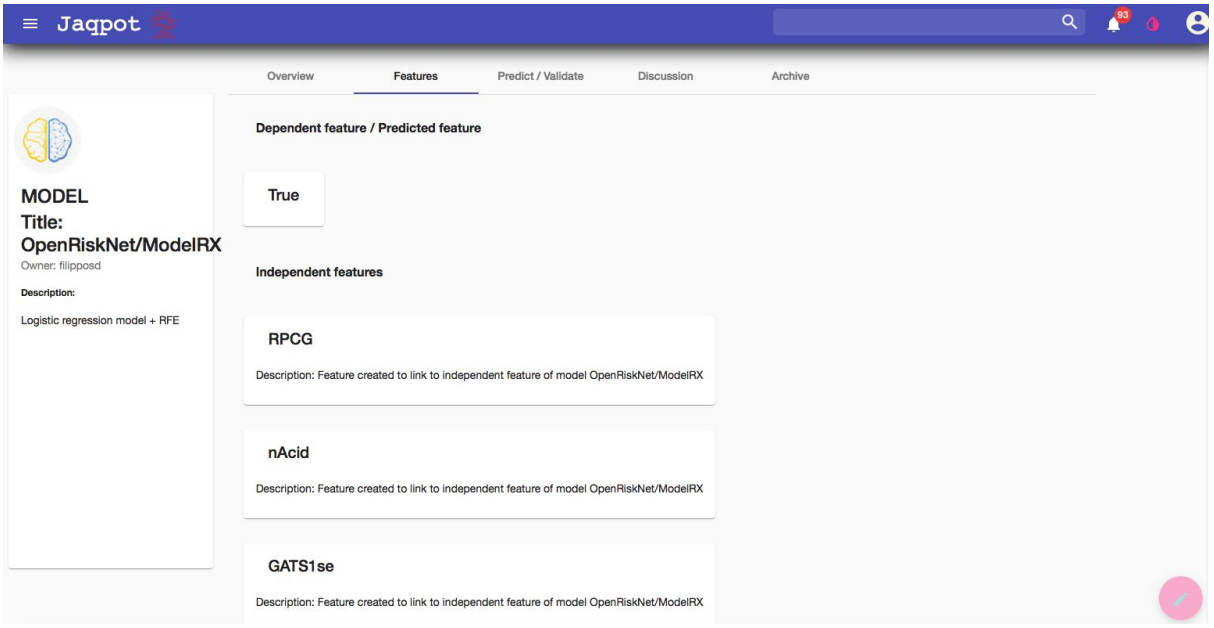
```
[35]: pred
```

[35]:

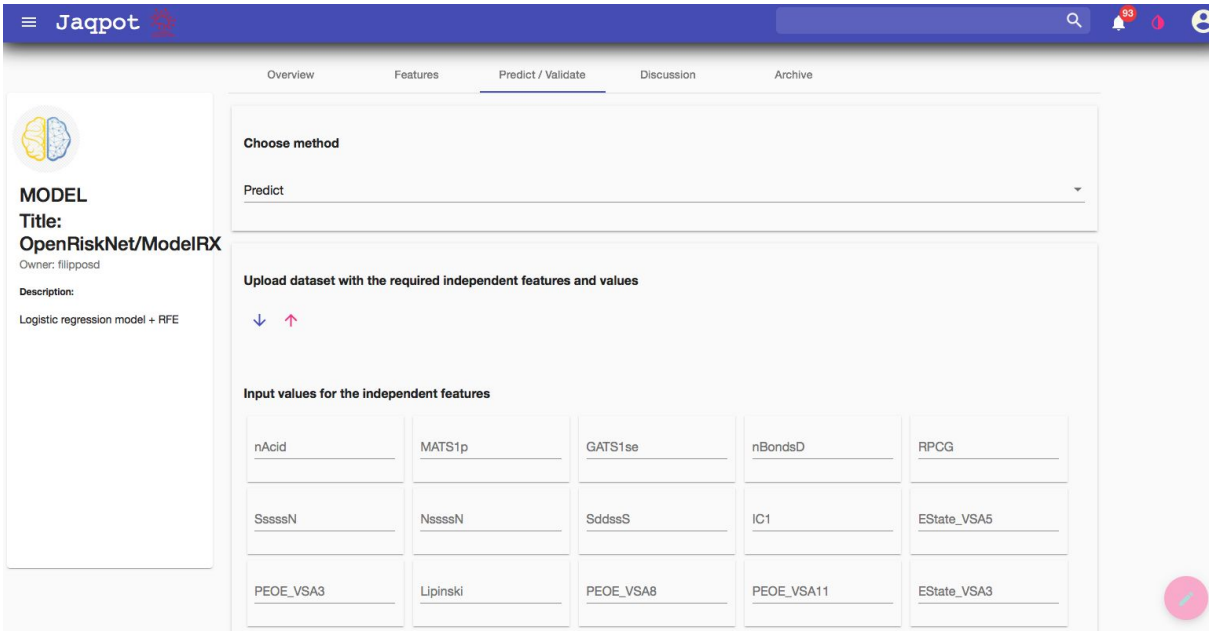
| | NssssN | PEOE_VSA3 | nAcid | SlogP_VSA10 | MATS1p | nBondsD | EState_VSA2 | GATS1se | SddssS | RPCG | ... | IC1 | Lipinski | VSA_EState8 | EState_VSA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0.177288 | 0.0 | 0.000000 | 0.525334 | 0.066667 | 0.274539 | 0.336161 | 1.000000 | 0.103911 | ... | 0.982863 | 1 | 0.217271 | 0.6187 |
| 1 | 0 | 0.177288 | 0.0 | 0.000000 | 0.535277 | 0.000000 | 0.000000 | 0.572574 | 1.000000 | 0.166547 | ... | 0.706634 | 1 | 0.199741 | 0.1463 |
| 2 | 0 | 0.177288 | 0.0 | 0.000000 | 0.535277 | 0.000000 | 0.000000 | 0.572574 | 1.000000 | 0.166547 | ... | 0.706634 | 1 | 0.199741 | 0.1463 |
| 3 | 0 | 0.000000 | 0.0 | 0.185059 | 0.554742 | 0.066667 | 0.000000 | 0.515808 | 1.000000 | 0.234713 | ... | 0.677722 | 1 | 0.708066 | 0.1286 |
| 4 | 0 | 0.333461 | 0.0 | 0.142857 | 0.542120 | 0.066667 | 0.129238 | 0.198597 | 1.000000 | 0.124416 | ... | 0.806137 | 1 | 0.402696 | 0.8211 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | ... | |
| 383 | 0 | 0.000000 | 0.0 | 0.000000 | 0.654844 | 0.000000 | 0.000000 | 0.487902 | 1.000000 | 0.287190 | ... | 0.682089 | 1 | 0.419335 | 0.4023 |
| 384 | 0 | 0.476721 | 0.0 | 0.189304 | 0.512897 | 0.266667 | 0.307252 | 0.215919 | 0.513168 | 0.154092 | ... | 0.955968 | 1 | 0.169934 | 0.0000 |
| 385 | 0 | 0.000000 | 0.0 | 0.000000 | 0.643280 | 0.000000 | 0.128742 | 0.452737 | 1.000000 | 0.099241 | ... | 0.599099 | 0 | 0.604214 | 0.0000 |
| 386 | 0 | 0.000000 | 0.0 | 0.000000 | 0.725519 | 0.133333 | 0.192371 | 0.246524 | 1.000000 | 0.210130 | ... | 0.508553 | 1 | 0.386176 | 0.2617 |
| 387 | 0 | 0.000000 | 0.0 | 0.000000 | 0.506561 | 0.066667 | 0.067106 | 0.155743 | 1.000000 | 0.186433 | ... | 0.591139 | 1 | 0.278014 | 0.6385 |

388 rows × 21 columns

Please note that the model is now also available over the Jaqpot user interface at https://ui-jaqpot.prod.openrisknet.org for users to inspect:




And also make predictions on their own data, either typed in or provided over a CSV file, using the auto-generated template provided by Jaqpot:




A more detailed presentation of Jaqpot's offering was presented during the webinar "Demonstration on OpenRiskNet approach on modelling for prediction or read across", available at https://openrisknet.org/events/67/, where both the recording and the slides are available.

The documentation for the Jaqpot 5 API has been made available over Swagger at https://api-Jaqpot.prod.openrisknet.org/Jaqpot/swagger/.

# Lazar

*lazar* (Lazy Structure-Activity Relationships developed at JGU/IST) is a framework based on Ruby libraries. It also depends on a couple of external programs and libraries. All required libraries will be installed with the gem install lazar command. To build this prediction model in lazar following steps where required. Executing the following commands either from an interactive Ruby shell or a Ruby script.

1. **Create the training dataset**

Create a CSV file with two columns. The first line should contain either SMILES or InChI (first column) and the endpoint (second column). The first column should contain either the SMILES or InChI of the training compounds, the second column the training compounds toxic activities (qualitative or quantitative). Add metadata to a JSON file with the same basename containing the fields "species", "endpoint", "source".

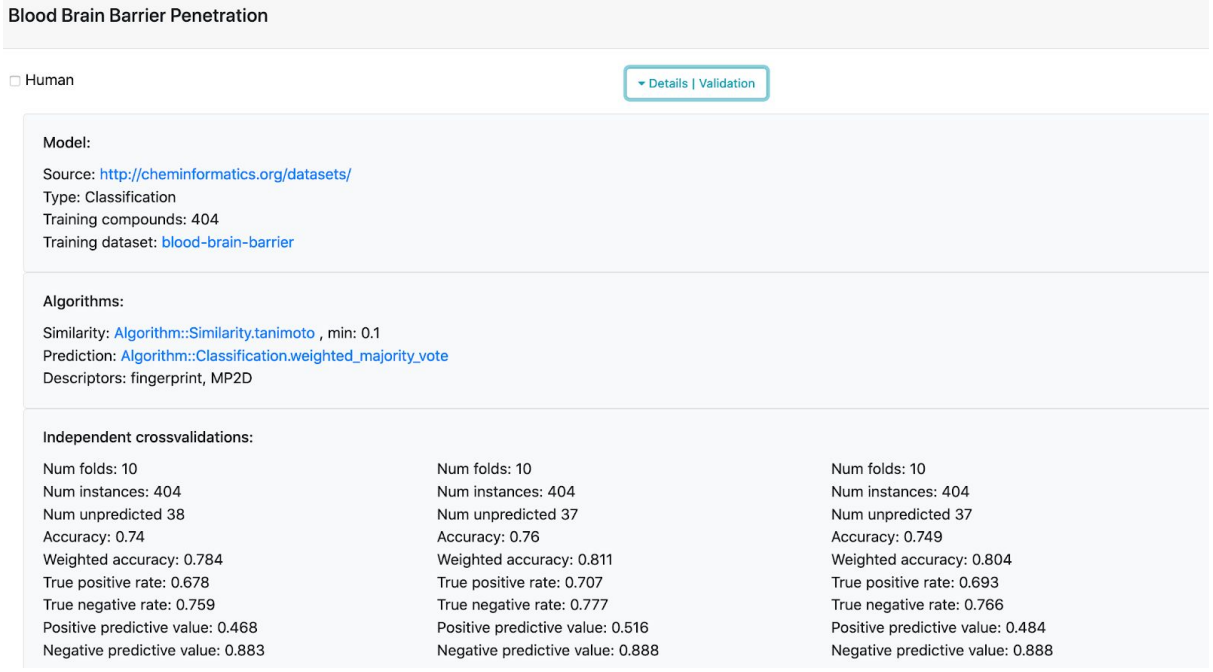
```
training_dataset = Dataset.from_csv_file "Blood_Brain_Barrier_Penetration-Human.csv"
```

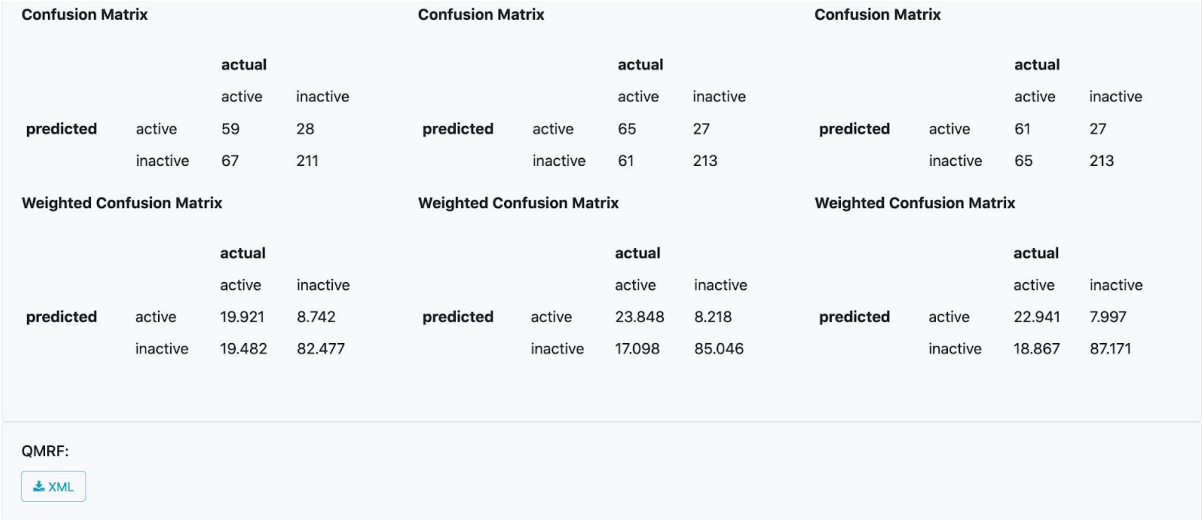
2. **Create and validate the lazar model with default algorithms and parameters**

```
validated_model = Model::Validation.create_from_csv_file Blood_Brain_Barrier_Penetration-Human.csv
```

This command will create a *lazar* model and validate it with three independent 10-fold cross validations.

Following screenshots represents the model details and validation results as shown in the *lazar* GUI (https://lazar.prod.openrisknet.org/predict).





**Confusion Matrix**

| | | actual | |
|---|---|---|---|
| | | active | inactive |
| predicted | active | 59 | 28 |
| | inactive | 67 | 211 |

**Confusion Matrix**

| | | actual | |
|---|---|---|---|
| | | active | inactive |
| predicted | active | 65 | 27 |
| | inactive | 61 | 213 |

**Confusion Matrix**

| | | actual | |
|---|---|---|---|
| | | active | inactive |
| predicted | active | 61 | 27 |
| | inactive | 65 | 213 |

**Weighted Confusion Matrix**

| | | actual | |
|---|---|---|---|
| | | active | inactive |
| predicted | active | 19.921 | 8.742 |
| | inactive | 19.482 | 82.477 |

**Weighted Confusion Matrix**

| | | actual | |
|---|---|---|---|
| | | active | inactive |
| predicted | active | 23.848 | 8.218 |
| | inactive | 17.098 | 85.046 |

**Weighted Confusion Matrix**

| | | actual | |
|---|---|---|---|
| | | active | inactive |
| predicted | active | 22.941 | 7.997 |
| | inactive | 18.867 | 87.171 |

QMRF:

XML

### Experiment with other algorithms

You can pass algorithm specifications as parameters to the Model::Validation.create_from_csv_file command. Algorithms for descriptors, similarity calculations, feature_selection and local models are specified in the algorithm parameter. Unspecified algorithms and parameters are substituted by default values.

The example below selects

MP2D fingerprint descriptors

Tanimoto similarity with a threshold of 0.1

no feature selection

weighted majority vote predictions

```
algorithms = {
:descriptors => { # descriptor algorithm
  :method => "fingerprint", # fingerprint descriptors
  :type => "MP2D" # fingerprint type, e.g. FP4, MACCS
},
:similarity => { # similarity algorithm
  :method => "Algorithm::Similarity.tanimoto",
  :min => 0.1 # similarity threshold for neighbors
},
:feature_selection => nil, # no feature selection
:prediction => { # local modelling algorithm
  :method => "Algorithm::Classification.weighted_majority_vote",
},
}
training_dataset = Dataset.from_csv_file "Blood_Brain_Barrier_Penetration-Human.csv"
model = Model::Validation.create  training_dataset: training_dataset, algorithms: algorithms
```

*lazar* is implemented as a RESTful service as well as a graphical user interface.

REST API: https://lazar.prod.openrisknet.org

GUI: https://lazar.prod.openrisknet.org/predict

# WEKA

In this section, we present a sample use case for employing the JGU Weka REST API for creating a predictive model based on a user-provided dataset. The model can be evaluated based on the already provided dataset or the user can use the generated model for evaluating a different dataset split kept as a test set. The web version of the JGU Weka REST API can be explored at the URL https://jguweka.prod.openrisknet.org/.

The BBB penetration dataset contains SMILES representations of chemical compounds and whether the particular compound can cross the blood-brain-barrier. In order to train a model for predicting the blood-brain-barrier penetrating or non-penetrating chemicals, we need some features for the chemical compounds, based on which we can train a model. The user can create the feature based dataset compatible with Weka ARFF format in any desired way. It is possible to use existing chemoinformatics libraries, e.g. CDK, Mordred, RDKit, etc. for extracting features from the chemical compounds, however, since the other services in this case study already use one or the other mentioned libraries, we will use a graph mining based feature extraction algorithm for chemical compounds in order to use features which provide a different perspective to the problem.

The existing libraries use handcrafted features which have been identified by chemical experts due to certain properties of these chemical structures/features. It has been demonstrated that elaborate patterns can also be used to summarize ground features. Using patterns which summarize several ground features also has the potential of revealing latent information not present in any ground feature. The **La**tent **St**ructure **P**attern **M**ining (LAST-PM) algorithm by Maunz et al.[10] aims to extract ground features from chemical compounds based on embedding relationships between individual patterns while taking into consideration the frequency and/or correlation of the patterns.

The LAST-PM implementation depends on a number of libraries, therefore, the algorithm has been containerised and a fully functional Docker image has been made available at https://hub.docker.com/r/jguweka/chem_descriptor_miner. The feature extraction application takes as input two files, (i) an “smi” file containing SMILES formatted chemical compounds, and (ii) a “class” file with the target/class variable corresponding to each chemical compound in the SMILES file. All the processing steps are then handled by the containerised application and feature extraction, conversion of chemical features from graph data format into SMARTS notation, and finally, the creation of a Weka ARFF file with the extracted features is carried out by the containerised application.

Assuming that we have separated the SMILES formatted chemical compounds and the target variable of the BBB dataset into an *smi* file and a *class* file, respectively, we can initiate the LAST-PM based feature extraction process using the following command

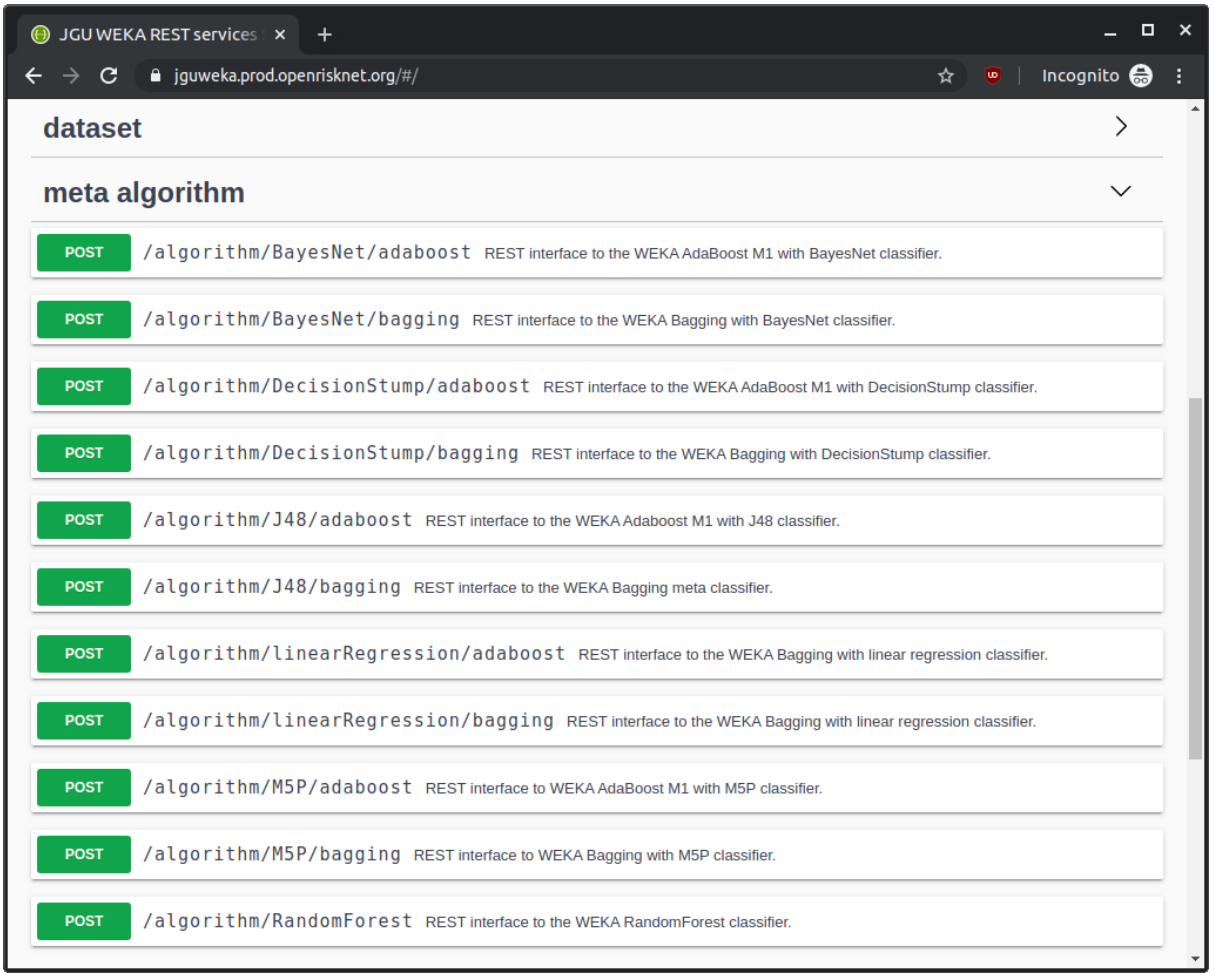
[10] Maunz A., Helma C., Cramer T., Kramer S. (2010) Latent Structure Pattern Mining. In: Balcázar J.L., Bonchi F., Gionis A., Sebag M. (eds) Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2010. Lecture Notes in Computer Science, vol 6322. Springer, Berlin, Heidelberg

to start the chem_descriptor_miner Docker container.

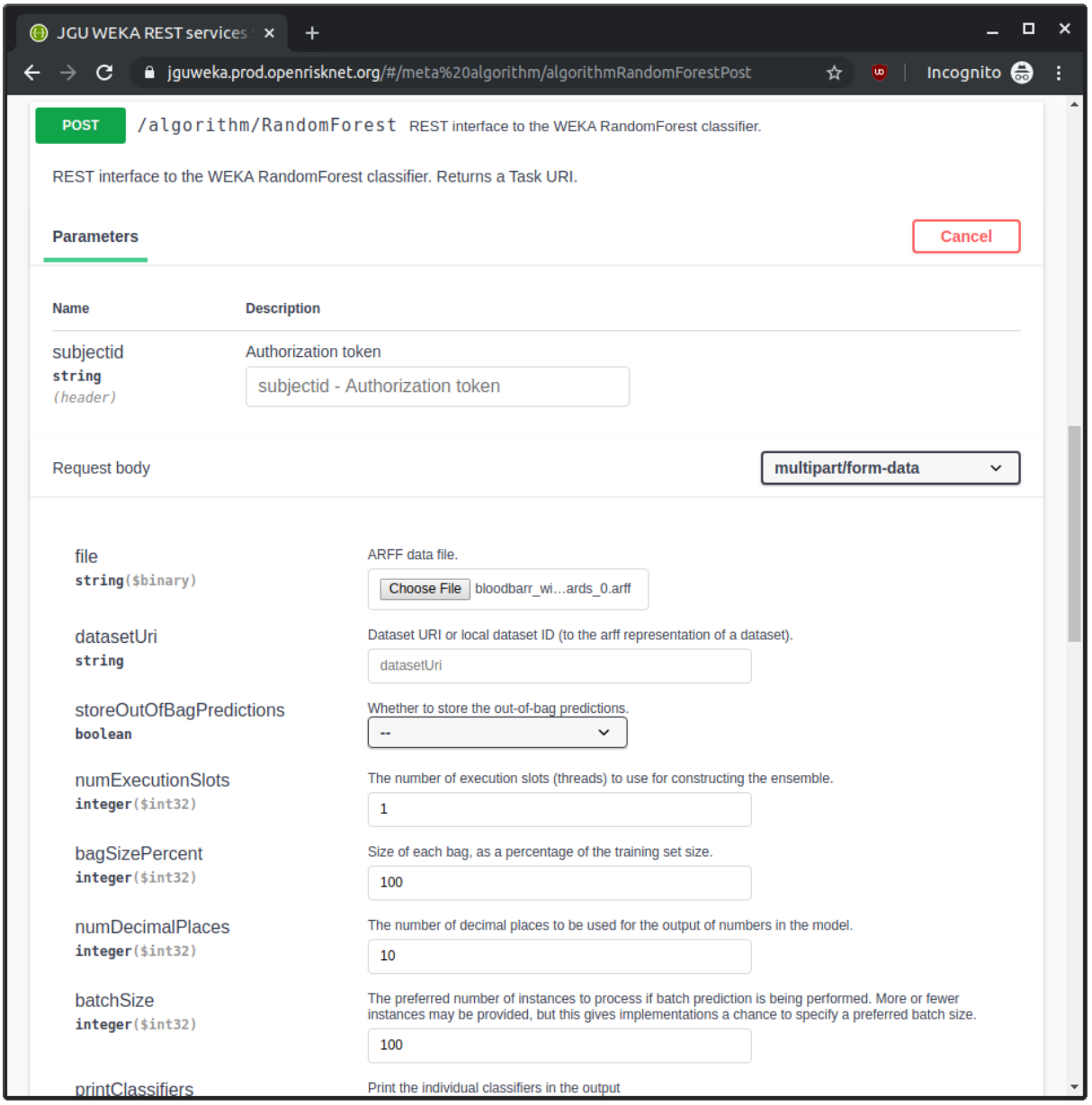
```
docker run -ti -v /path/to/data_dir:/work jguweka/chem_descriptor_miner /bin/bash
```

Here, the `/path/to/data_dir` is the host computer's directory containing the smi and class files. The resulting ARFF files are also saved to the same directory on the host machine. Running the above command starts the Docker container, which contains the main script in the `chem-descriptors` directory and the smi and class files are linked from the `/path/to/data_dir` directory in the host machine to the `work` directory in the Docker container. The different program options, e.g. dataset name, minimum descriptor frequency, number of ground features, SMARTS wildcarding and aromatic annotations, etc. can be adjusted by editing the `script.sh` BASH script in the `chem-descriptors` directory before running the task of feature extraction. Once the process completes, the ARFF file is created in the host computer's data directory `/path/to/data_dir`.

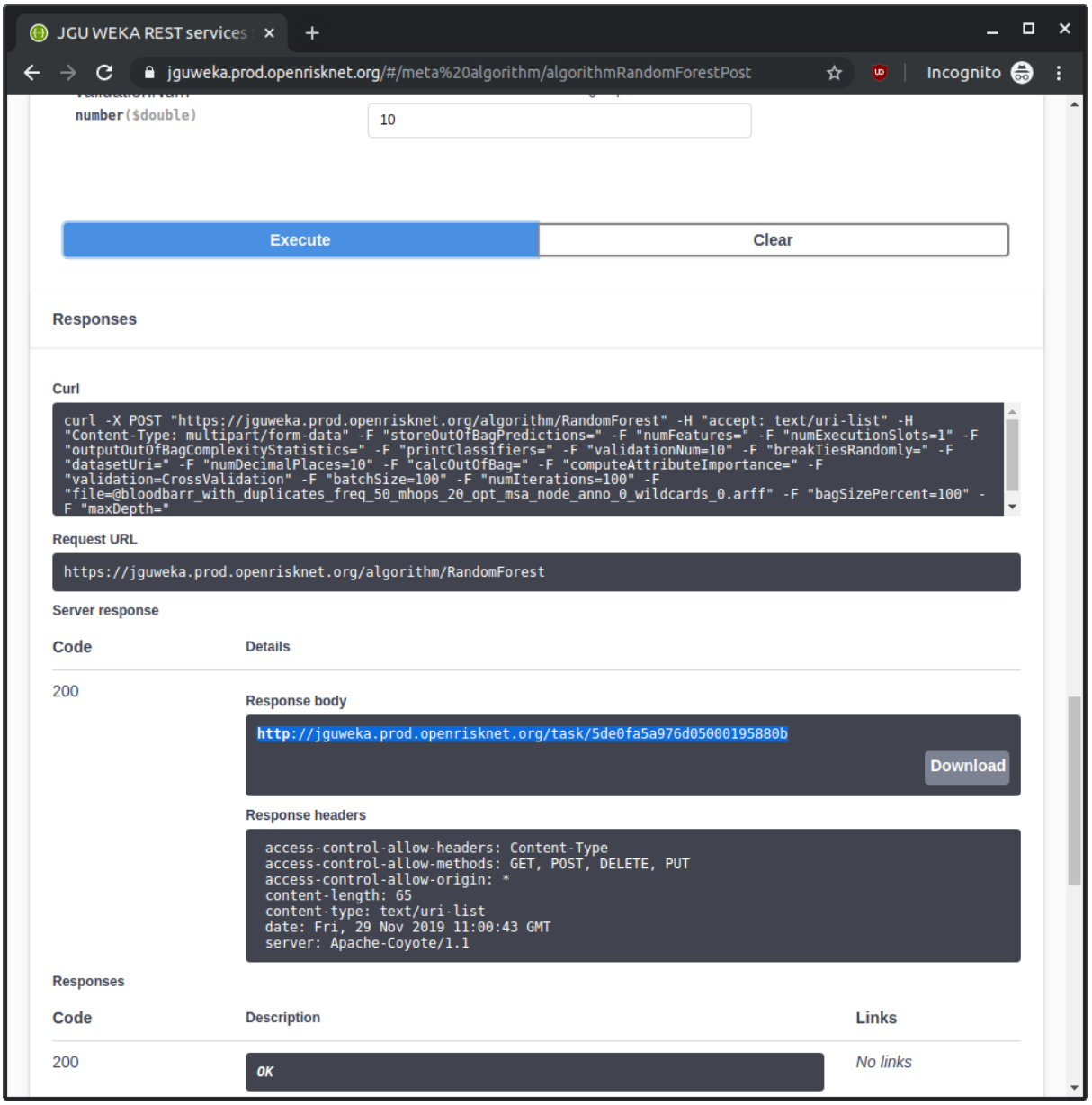
Here we assume that the BBB dataset has been successfully processed by the `jguweka/chem_descriptor_miner` containerised application. The Weka REST API exposes a number of machine learning algorithms. We will use the Random Forest algorithm for this demonstration using the default parameters. Apart from appearing under the *algorithm* category, the ensemble methods are also grouped under the *meta algorithm* category.



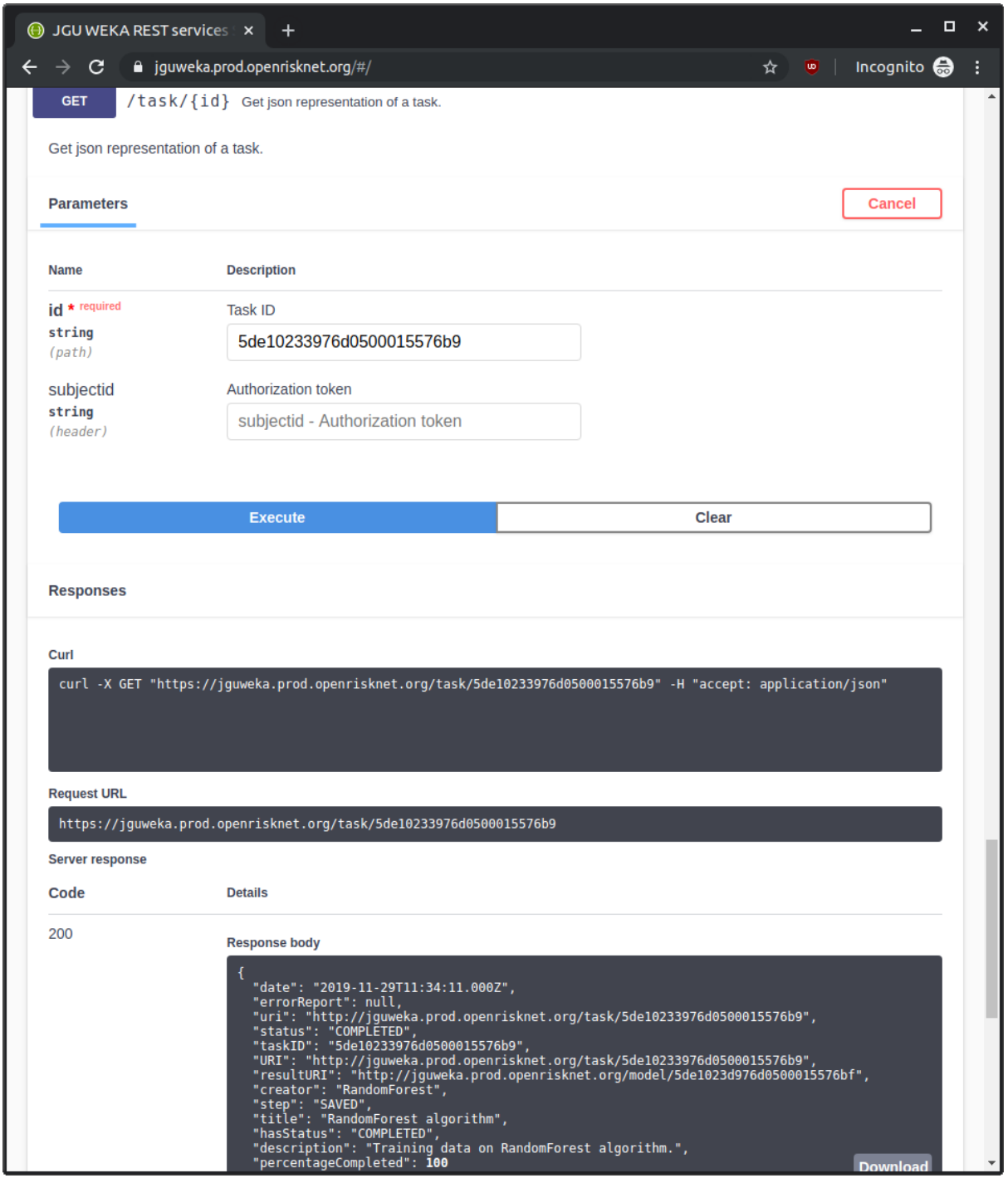

Selecting an algorithm entry opens the details for the particular algorithm. Clicking the “Try it out” button allows the user to upload a dataset or provide a URI and run the algorithm for the given data. The interface provides default values for the different parameters required for the selected algorithm. Here, we are using the default values.


JGU WEKA REST services
jguweka.prod.openrisknet.org/#/meta%20algorithm/algorithmRandomForestPost
Incognito
POST
/algorithm/RandomForest
REST interface to the WEKA RandomForest classifier.
REST interface to the WEKA RandomForest classifier. Returns a Task URI.
Parameters
Cancel
Name
Description
subjectid
string
(header)
Authorization token
subjectid - Authorization token
Request body
multipart/form-data
file
string($binary)
ARFF data file.
Choose File
bloodbarr_wi...ards_0.arff
datasetUri
string
Dataset URI or local dataset ID (to the arff representation of a dataset).
datasetUri
storeOutOfBagPredictions
boolean
Whether to store the out-of-bag predictions.
--
numExecutionSlots
integer($int32)
The number of execution slots (threads) to use for constructing the ensemble.
1
bagSizePercent
integer($int32)
Size of each bag, as a percentage of the training set size.
100
numDecimalPlaces
integer($int32)
The number of decimal places to be used for the output of numbers in the model.
10
batchSize
integer($int32)
The preferred number of instances to process if batch prediction is being performed. More or fewer instances may be provided, but this gives implementations a chance to specify a preferred batch size.
100
printClassifiers
Print the individual classifiers in the output

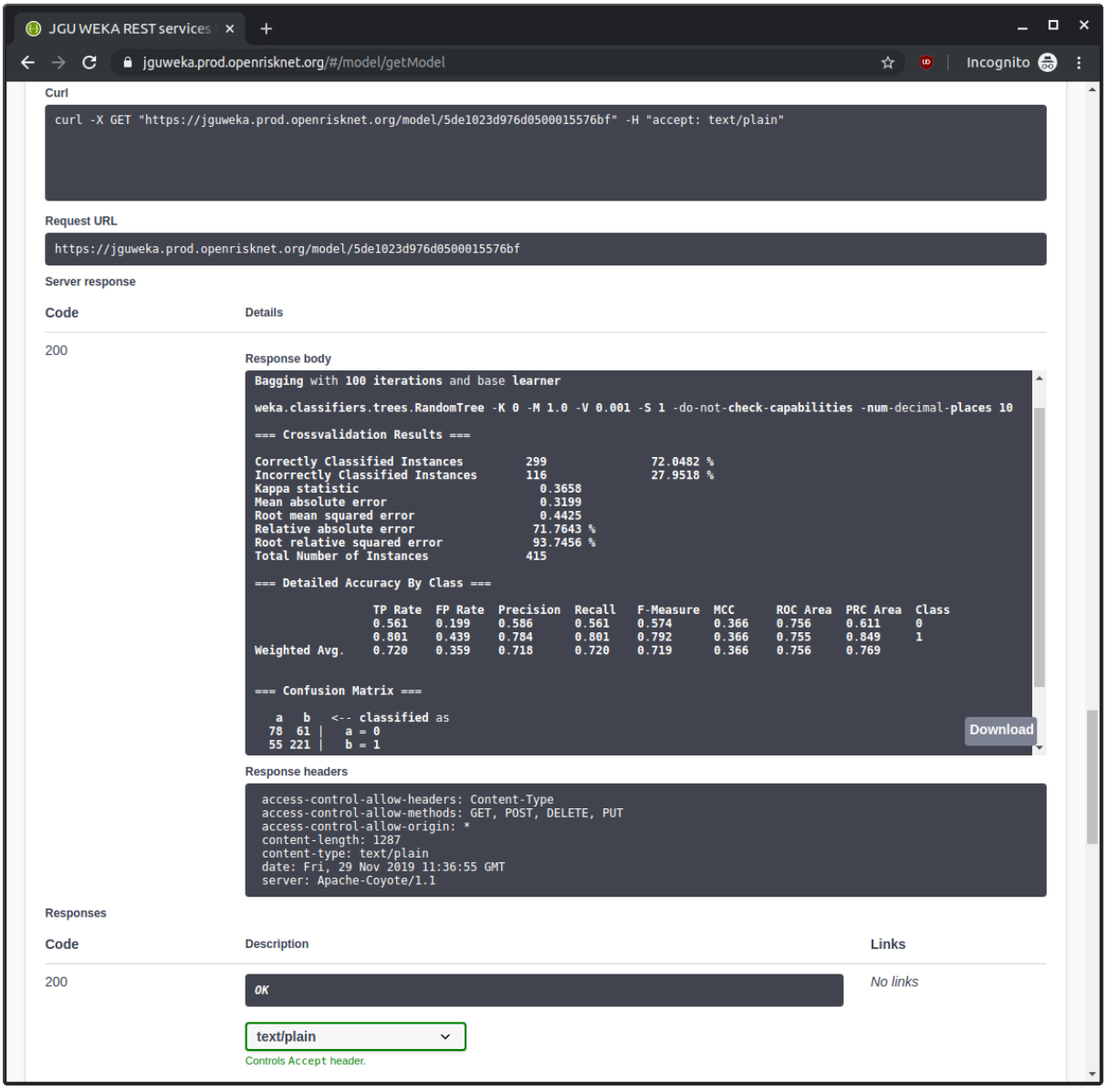
Clicking the “Execute” button creates a task for model creation. The task link is returned as a response to the execute command.



The task ID can be used to query the server about the state of the modelling task.



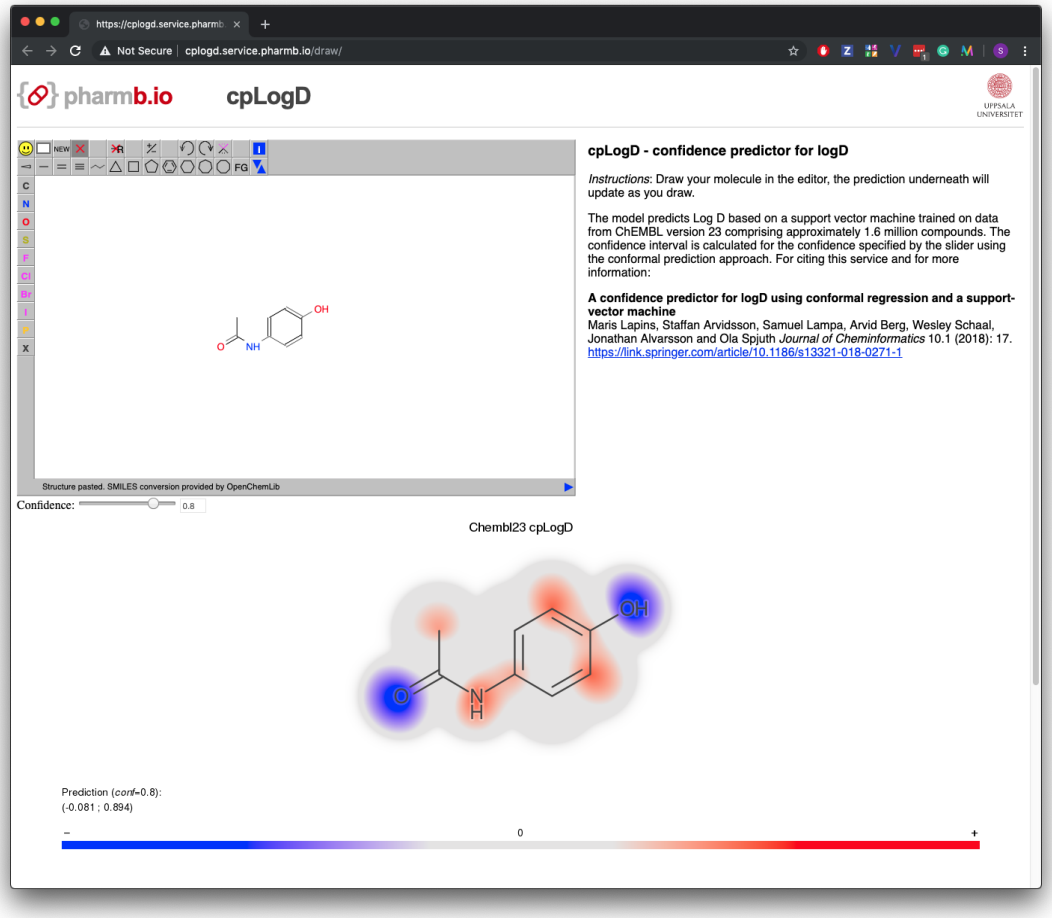
Once the server returns the COMPLETED status for the task, the model can be retrieved based on the model ID given under resultURI. The generated model is presented in human readable form but it can also be downloaded as a JSON object.




In this case, the generated model was able to achieve a 72% accuracy. The outlined steps can also be carried out using REST calls through a Jupyter Notebook or in any other scripts the user is writing for their predictive task. The REST calls for each step are also shown in the Swagger UI based front end.

## CPSign

CPSign is a licensed software co-developed by the UU team. This software brings together Conformal Prediction and cheminformatics, accessible through a Java API, command line interface and a Web UI. Currently it supports loading chemistry from various formats, basic filtration of data, descriptor generation using the Signatures descriptor and predictive modeling using Transductive (TCP) and Inductive conformal prediction (ACP and CCP) as well as Cross Venn-ABERS prediction (CVAP). Models can be trained from a web UI as well as a OpenAPI documented REST API. Currently the REST API supports uploading of license files, datasets and training of Venn-ABERS based classification models. Since CPSign is a licensed software the functionality requires authentication using Keycloak. The web UI is located at the URL: http://modelingweb.prod.openrisknet.org/ and the Swagger UI for the OpenAPI definition is available at the URL: http://modelingweb.prod.openrisknet.org/swagger-ui/. Trained models can, aside from previous modes of access, be deployed as microservices in OpenShift and expose a REST API described using OpenAPI. Each microservice also include a GUI where users can load molecules or draw them on their own, continuously making predictions as the molecule is edited in the GUI. An example of the drawing GUI is shown below, for a model predicting the LogD value. Colouring of the atoms show how individual atoms contribute to the prediction (blue contribute towards a lower LogD and red towards an increased LogD).


https://cplogd.service.pharmb
Not Secure
cplogd.service.pharmb.io/draw/
pharmb.io
cpLogD
UPPSALA UNIVERSITET
cpLogD - confidence predictor for logD
Instructions: Draw your molecule in the editor, the prediction underneath will update as you draw.
The model predicts Log D based on a support vector machine trained on data from ChEMBL version 23 comprising approximately 1.6 million compounds. The confidence interval is calculated for the confidence specified by the slider using the conformal prediction approach. For citing this service and for more information:
A confidence predictor for logD using conformal regression and a support-vector machine
Maris Lapins, Staffan Arvidsson, Samuel Lampa, Arvid Berg, Wesley Schaal, Jonathan Alvarsson and Ola Spjuth Journal of Cheminformatics 10.1 (2018): 17.
https://link.springer.com/article/10.1186/s13321-018-0271-1
Structure pasted. SMILES conversion provided by OpenChemLib
Confidence:
0.8
Chembl23 cpLogD
OH
O
NH
Prediction (conf=0.8):
(-0.081 ; 0.894)
–
0
+

OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

# Case Study

## A systems biology approach for grouping compounds [**SysGroup**]

# SUMMARY

This case study will use the approach of the diXa / DECO2 (Cefic-LRI AIMT4[1]) projects to reproduce and extend the results obtained on the identification of hepatotoxicant groups based on similarity in mechanisms of action (omics-based) and chemical structure using services from OpenRiskNet. Figure 1 displays the workflow that was used in this case study.

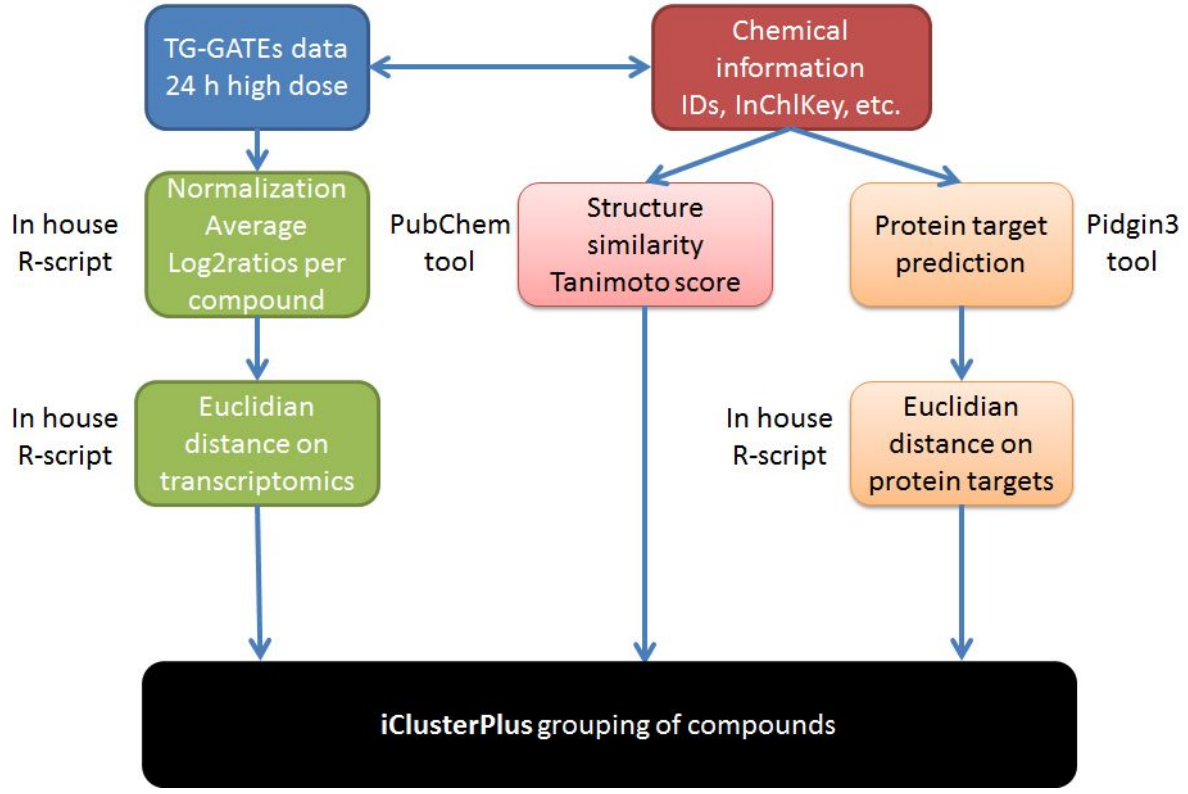



**Figure 1**. Workflow for grouping compounds by integrating transcriptomics data, Tanimoto similarity scores and ligand-based target predictions using iClusterPlus.

[1] http://cefic-lri.org/projects/aimt4-um-deco2-moving-from-deco-towards-oecd/

# DESCRIPTION

## Implementation team

Coordination:

- Danyel Jennen, Maastricht University, Department of Toxicogenomics

Other members:

- Jumamurat Bayjanov, Maastricht University, Department of Toxicogenomics

## Case Study objective

The objective of this case study is to implement an integrated analysis using chemoinformatics and omics data for improved grouping of compounds with similar toxicity and/or mode of action.

## Risk assessment framework

SysGroup covers the identification of use scenario / chemical of concern / collection of existing information (Tier 0 in the selected framework) and its steps related to:

- Identification of molecular structure;
- Collection of support data;
- Identification of analogues / suitability assessment and existing data.

# DEVELOPMENT

## Databases and tools

In this case study TG-GATEs transcriptomics data from primary human hepatocytes exposed to 139 compounds for 24h and the highest dosage were used. This dataset was obtained from the OpenRiskNet service "Transcriptomics data from human, mouse, rat in vitro liver models"[2].

For the 139 compounds PubChem[3] was used to obtain 2D Tanimoto scores and PIDGIN3[4] was used to retrieve ligand-based target predictions.

Integration of the transcriptomics data with the chemoinformatics data was performed using iClusterPlus[5], an integrative clustering framework developed to integrate diverse data types (i.e. binary, categorical, and continuous values) by a latent variable approach.

## Technical implementation

Integration with other case studies is needed. SysGroup acquires information and data from the DataCure or TGX case study and can feed into AOPLink and ModelRX.

Through the services ToxPlanet and ToxicoDB of the Implementation Challenge winners Toxplanet and UHH, respectively, information on the obtained groups of chemicals is obtained.

Currently available services:

- Transcriptomics data from human, mouse, rat in vitro liver models
    - Repository for transcriptomics data from multiple in vitro human, rat and mouse toxicogenomics projects
    - Service type: Database / data source

[2] https://openrisknet.org/e-infrastructure/services/164/
[3] https://pubchem.ncbi.nlm.nih.gov/score_matrix/score_matrix.cgi
[4] https://pidginv3.readthedocs.io/en/latest/
[5] https://bioconductor.org/packages/release/bioc/html/iClusterPlus.html

# OUTCOMES

The outcome of this case study is the workflow as shown in Figure 1. This workflow was setup using GNU Make[6] and is available through OpenRiskNet's GitHub[7].

The workflow contains several steps to integrate 3 different datasets or data types for the purpose of grouping chemicals based on similarity in mechanisms of action (omics-based) and their chemoinformatics properties.

**Step 1**: Obtaining transcriptomics data

- Transcriptomics data were obtained from the OpenRiskNet service "Transcriptomics data from human, mouse, rat in vitro liver models"[8]. Here we only selected normalized data of from TG-GATEs (i.e. from primary human hepatocytes exposed to 139 compounds for 24h and the highest dosage);
- Scaled Euclidean distances (scale 0-1; 0 = most similar;1 = most dissimilar) were calculated from these transcriptomics profiles;

→ Result step 1: a 139 x 139 distance matrix for transcriptomics data with ChEMBL ID as identifier.

**Step 2**: Calculating a Tanimoto score for each compound

- Convert ChEMBL ID to InChIKey using ChEMBL conversion tool and subsequently InChIKey to CID using PubChem conversion tool;
- Use the list of CID's to calculate 2D Tanimoto scores via PubChem tool, which are shown as percentages;
- Convert percentages to 0-1 scale;
- Convert CID to ChEMBL ID;
- Convert 2D Tanimoto scores to a distance, by subtracting Tanimoto score from 1;

→ Result step 2: a 139 x 139 distance matrix for 2D Tanimoto scores with ChEMBL ID as identifier.

**Step 3**: Predicting ligand-based protein targets for each compound

- Convert ChEMBL ID to smiles ChEMBL conversion tool;
- Predict ligand-based protein targets from smiles using Pidgin3 tool;
- Scaled Euclidean distances (scale 0-1; 0 = most similar;1 = most dissimilar) were calculated from these protein target profiles;
- Convert smiles to ChEMBL ID;

→ Result step 3: a 139 x 139 distance matrix for ligand-based protein targets with ChEMBL ID as identifier.

**Step 4**: Grouping of chemicals from integrated data

- For the obtained matrices from step 1-3 the order of the ChEMBL ID's are set in the same order;
- The 3 matrices are integrated using iClusterPlus; the number of clusters is set to 46 (~1/3 of 139);

→ Result step 4: heatmaps for each data type sorted per cluster (see Figure 2).

[6] https://www.gnu.org/software/make/
[7] https://github.com/OpenRiskNet/notebooks/tree/master/openrisknet_sysgroup
[8] https://openrisknet.org/e-infrastructure/services/164/

The grouping of the chemicals in these clusters needs to be further investigated using toxicological data from the DataCure case study and various OpenRiskNet services, such as **ToxicoDB** and **ToxPlanet**. However, due to time constraints this final step was not achieved within the time frame of the project.

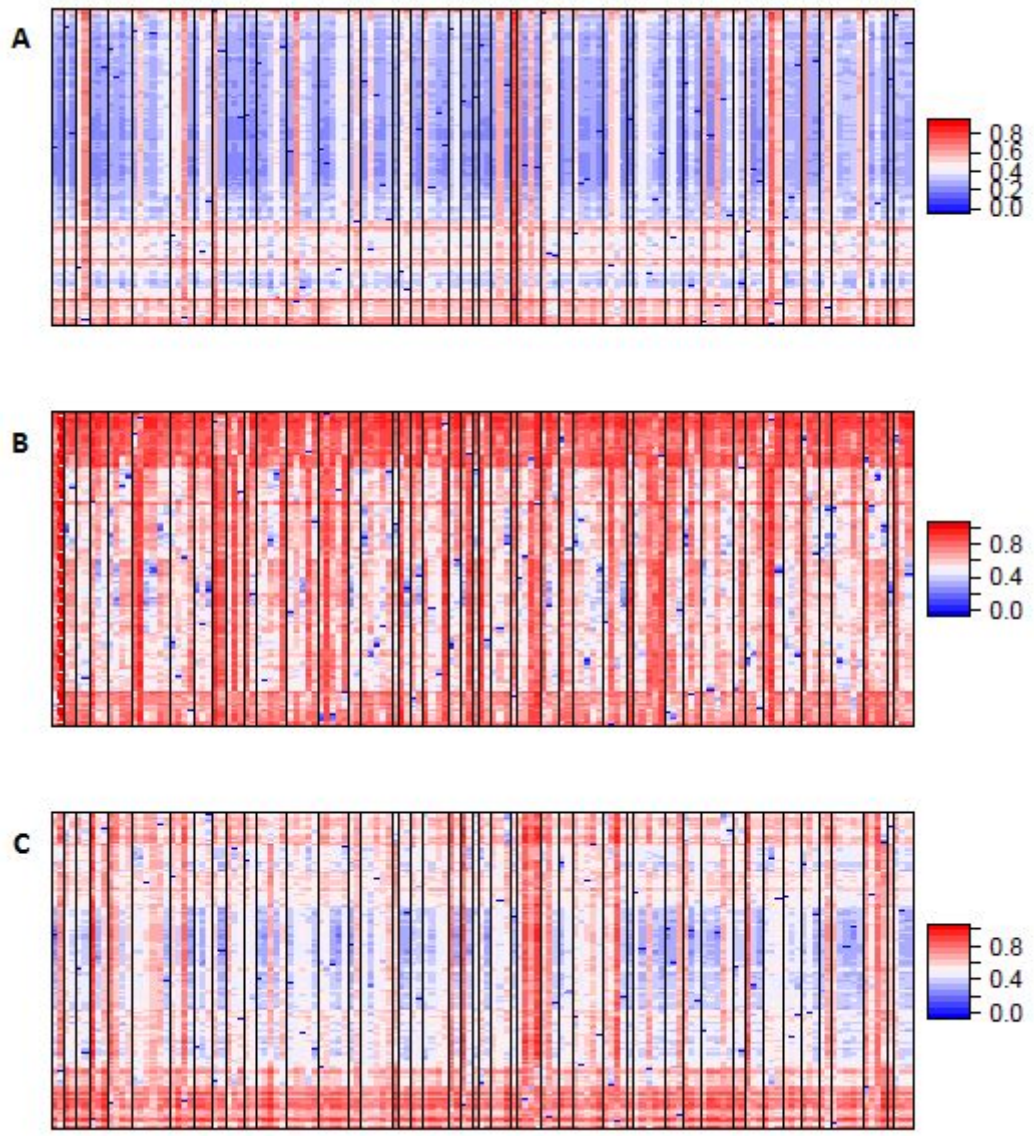



**Figure 2**. iClusterPlus result showing the grouping of 139 compounds in 46 clusters; A) transcriptomics data, B) 2D Tanimoto scores, and C) ligand-based protein predictions

# Related resources

**OpenRiskNet Part II: Predictive Toxicology based on Adverse Outcome Pathways and Biological Pathway Analysis**

Marvin Martens, Thomas Exner, Nofisat Oki, Danyel Jennen, Jumamurat Bayjanov, Chris Evelo, Tim Dudgeon, Egon Willighagen

28 August 2019 | Poster

OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

# Case Study

## Metabolism Prediction [**MetaP**]

# SUMMARY

Metabolites may well play an important role in adverse effects of parent drug or other xenobiotic compounds. In this case study VU (CS leader), HITeC/HHU (associate partner and implementation challenge winner), JGU, and UU have worked together on making methods and tools available for metabolite and site-of-metabolism (SOM) prediction. For that purpose we integrated and used ligand-based metabolism predictors (e.g. MetPred, enviPath, FAME, SMARTCyp) and we incorporated protein-structure and -dynamics based approaches to predict SOMs by Cytochrome P450 enzymes (P450s). P450s metabolise ~75% of the currently marketed drugs and their active-site shape and plasticity often play an important role in determining the substrate's SOM. It is expected that this work will be continued after the end of the project to make services available for the prediction of microbial biotransformation pathways by integrating the enviPath data and software developed in part by JGU.

During method development, model calibration and validation we used databases such as XMetDB and other open-access databases for drugs, xenobiotics and their respective metabolites. To facilitate the combined use of the metabolite prediction approaches and their outcomes, we benefited of ongoing development in workflow management systems and we made Jupyter Notebooks available to facilitate collection and visualization of results from the different available services. We illustrated the added value of having multiple predictors and our Jupyter notebooks available, in a pilot study on retrospective consensus predictions of known SOMs for drug compounds for which possible metabolite-associated toxicity was previously reported.

# DESCRIPTION

## Implementation team

| CS leader | Team |
|---|---|
| Daan Geerke (VU) | VU, UU, JGU, HITeC/HHU |

## Case Study objective

The objective of this case study was to enable and facilitate metabolite prediction within the OpenRiskNet infrastructure and to evaluate and demonstrate the added value of it. For that purpose we integrated different tools for metabolism prediction, including tools for:

- Ligand-based site-of-metabolism (SOM) prediction using reaction SMARTs, circular fingerprints and/or atomic reactivities;
- QSBR (quantitative-structure biotransformation relationship) modeling of microbial biotransformation;
- Protein-structure and -dynamics based prediction of CYP450 isoform specific binding and SOMs;
- Predicting probabilities for specific reaction type events.

Combined use of the tools has been made possible and compared using Jupyter notebooks that gather and visualize results from the available case-study services.

See the “Databases and tools” subsection for more details on the corresponding tools. For our comparisons of predictive (and consensus) performance we used selected compounds from literature for which SOMs and metabolite-associated toxicity have been reported. We anticipate to present our results in an upcoming manuscript on tool integration, which will illustrate how using several tools can have additional value (when compared to individual tools) to (site-of-)metabolism prediction.

## Risk assessment framework

Prediction outcomes can serve as input for other molecular structure-based AO predictors, which relates to Tier 0 (Step 1: identification of molecular structure) and Tier 1 (Step 6: mechanism of action).

# DEVELOPMENT

## Databases and tools

The table below gives an overview of metabolite prediction tools that are integrated and have been used in this case study. During method development, model calibration, and validation, advantage was taken of data from XMetDB (reference 1) and other databases for drugs, xenobiotics and their respective metabolites, as available in ZINC, ChEMBL, DrugBank, EAWAG-BBD and/or the SMARTCyp and FAME suites. Integration of enviPath (envipath.org) is still ongoing, which is a database and prediction system for microbial biotransformation of organic environmental contaminants.[2-4]

**Table 1:** Currently available MetaP tools.

| Tool | Input | Output | Method |
|---|---|---|---|
| MetPred (UU) | 2D chemical structure of ligand | SOMs with Reaction Types for Phase I reactions | Preprocess Metabolite reaction database (>100K biotransformations) using MCS. For each query compound, look up similar atom environments based on circular fingerprints and use ReactionSMARTS to identify reaction types. See metpred.service.pharmb.io/draw/ |
| FAME 3 (HHU/ HITeC) | 2D chemical structure of ligand | SOMs for Phase I, Phase II, or combined Phase I/II metabolism | Machine learning using 2D-circular-environment based atomic descriptors, see reference 5. |
| SMARTCyp 2.0 (external) | 2D chemical structure of ligand | Rank atoms (SOMs) for P450-isoform specific reactions | Combining reactivity (from database on QM calculated transition state energies) with simple 2D molecular accessibility descriptors for SOM prediction. See reference 6 and smartcyp.sund.ku.dk/mol_to_som |
| Plasticity tools (VU) | 3D Chemical structure of ligand | Prediction of most probable SOMs for P450-isoform specific reactions | Protein-structure and dynamics based prediction of substrate binding orientations and corresponding SOM in the active site of CYP isoforms (1A2, 2D6, 3A4). Cf. reference 7. |

## Technical implementation

As summarised in Table 1, several services have come available in the MetaP case study. The listed services offer their functionality through RESTful APIs that are formalised according to OpenAPI specifications. The APIs are build using the Swagger toolchain and subsequently enable direct user interaction with the API endpoints using a browser-based User Interface (the Swagger UI). In addition, MetPred and SMARTCyp offer a custom

browser-based interface to their service (see links in Table 1). The APIs enable access to the core features of the services as summarised above, and typically accept submissions of chemical structures in common file formats.

API endpoint input and output data exchange is standardised to a machine-readable JSON format. Together with the OpenAPI data type definitions and JSON-LD data annotation it ensures seamless integration of the containerised services in the OpenRiskNet infrastructure and data exchange with other services.

Service API use and interoperability of the listed services is demonstrated using a Jupyter Notebook freely available on Github. Single 3D ligand structures in Tripos MOL2 format are used as input to the various services and the standardised JSON output are aggregated into a Pandas DataFrame demonstrating interoperability. Predicted SOMs are visualized on the 2D ligand depiction using the RDKit package.

# OUTCOMES

In addition to the service integration of the metabolite prediction tools listed above, we have evaluated the added value of having multiple tools and their combined use available (via Jupyter Notebooks). The different predictors give complementary types of output, *cf.* Table 1. MetPred, FAME 3, and the VU and SMARTCyp tools predict SOMs related to Phase I, Phase I/II, and Cytochrome P450 isoform specific conversion, respectively. Per (heavy) atom, normalized propensities are written out to indicate the likelihood of the atom to be a SOM. In addition, MetPred also gives back most probable reaction types at predicted SOMs. Facilitated by the Jupyter Notebook that supplies and visualizes output from the different predictors (Appendix 1), the MetaP tools can thus aid experts in guiding decision making on metabolite formation and/or in obtaining input for subsequent case studies.

The added value of having the multiple complementary tools available for metabolite prediction is illustrated by the Jupyter-notebook output presented in Appendix 1, which collects SOM predictions and MetPred predictions of Phase I reaction types (and which color-highlights atoms as predicted SOM if propensities are larger than a preset cutoff) for the three compounds in Figure 1. These compounds were selected because possible toxicological effects have been related with their metabolites, and their metabolism is extensively studied in literature (see Figure 1 for the experimentally determined SOMs).[8-10]

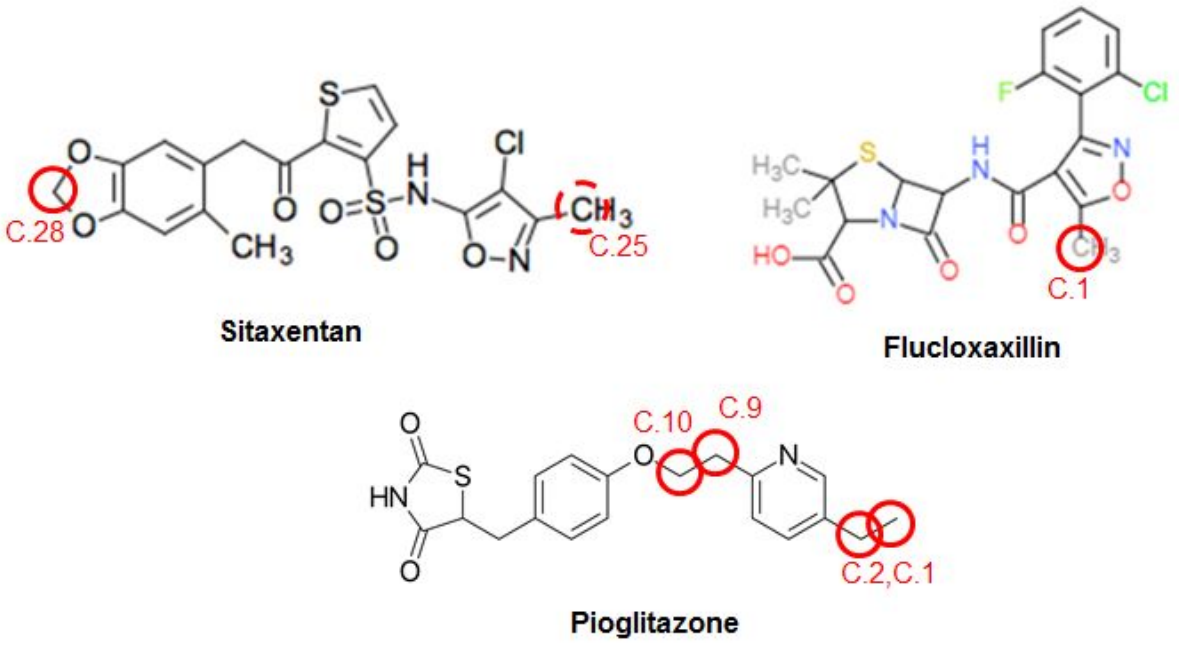



**Figure 1.** Molecular structures of Sitaxentan, Pioglitazone and Flucloxacillin, together with their experimentally determined sites-of-metabolism[8-10] as indicated by circles and atom indices.

Appendix 1 demonstrates that for Sitaxentan, consensus is obtained with the different tools in (correctly) predicting C.28 as SOM: FAME 3 and SMARTCyp 2.0 assign it as the most probable metabolic site, while docking confirms a possible reactive binding orientation in CYP (3A4). SMARTCyp also appoints C.25 as reactive and a possible SOM, which was identified in metabolism studies with dog liver microsomes.[8] In addition, the more mechanistic based SMARTCyp and docking tools can help in identifying the SOM-assignment by MetPred of O.17/18 as a false positive. Similarly, the zero scores of SMARTCyp and docking identify the predictions of MetPred and FAME 3 for C.23 of Pioglitazone to be a false positive as well, Appendix 1. For this compound, consensus is reached that C.9 and C.10 are possible sites of metabolism, while the current unavailability of a docking model for P450 2C8 may well partly explain why not all tools identify C.1 and C.2 both as possible SOM (in that case 2C8 is known to be the major involved P450 isoform[9]). It should also be noted that based on the significant scores for three out of four predictors, experts may (wrongly) assign Pioglitazone's S.25 as a possible SOM as well. As a third example, Appendix 1 illustrates the obtained consensus in correctly assigning C.1 of Flucloxacillin as the most probable metabolic site.

In conclusion, these examples illustrate how combining and comparing output from the different tools available in MetaP (and how collecting and visualizing their output in a Jupyter Notebook) can aid in and increase the value of SOM prediction when compared to having individual tools available alone.

# REFERENCES


1. O. Spjuth, P. Rydberg, E.L. Willighagen, C.T. Evelo, N. Jeliazkova, *J. Cheminf.* **2016**, *8*, 47.
2. J. Wicker, K. Fenner, L. Ellis, L. Wackett, S. Kramer, *Bioinformatics* **2010**, *26*, 814-821.
3. J. Wicker, T. Lorsbach, M. Gütlein, E. Schmid, D.A.R.S. Latino, S. Kramer, K. Fenner, *Nucleic Acids Res.* **2016**, *44*, D502-D508.
4. D.A.R.S. Latino, J. Wicker, M. Gütlein, E. Schmid, S. Kramer, K. Fennerad, *Environ. Sci.: Processes Impacts* **2017**, *19*, 449-464
5. M. Sicho, C. Stork, A. Mazzolari, C. de Bruyn Kops, A. Pedretti, B. Testa, G. Vistoli, D. Svozil, J. Kirchmair, *J. Chem. Inf. Model.* **2019**, *59*, 3400-3412.
6. P. Rydberg, D.E. Gloriam, J. Zaretzki, C. Breneman, L. Olsen, *ACS Med. Chem. Lett.* **2010**, *1*, 96-100.
7. J. Hritz, A. de Ruiter, C. Oostenbrink, *J. Med. Chem.* **2008**, *51*, 7469-7477.
8. J.C.L. Erve, S. Gauby, J.W. Maynard Jr., M.A. Svensson, G. Tonn, K.P. Quinn, *Chem. Res. Toxicol.* **2013**, *26*, 926-936.
9. T. Jaakkola, J. Laitila, P.J. Neuvonen, J.T. Backman, *Basic Clin. Pharmacol. Toxicol.* **2006**, *99*, 44-51.
10. R.E. Jenkins, X. Meng, V.L. Elliott, N.R. Kitteringham, M. Pirmohamed, B.K. Park, *Proteomics Clinic. Applic.* **2009**, *3*, 720-729.

# APPENDIX

## Example Jupyter notebook and predicted SOMs

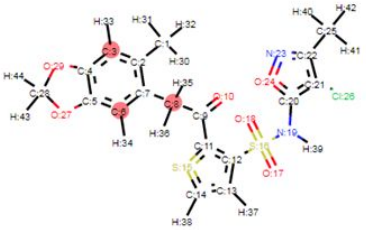
In [1]:
```
%%javascript
IPython.OutputArea.prototype._should_scroll = function(lines) {
    return false;
}
```

In [2]:
```
import os
import requests
import pandas

from rdkit import Chem
from rdkit.Chem.Draw import rdMolDraw2D
from IPython.display import HTML
from example_helpers import *

MDSTUDIO_URL = 'http://mdstudio-smartcyp.dev.openrisknet.org/'
FAME_URL = 'http://fame3.dev.openrisknet.org'
METPRED_URL = 'http://metpred.prod.openrisknet.org/v2'

data = get_dataset()
```

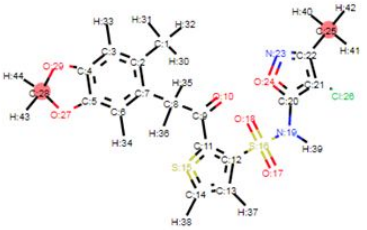
**SOM prediction comparing Docking, SMARTCyp, FAME and MetPred**

In [3]:
```
selection = ('sitaxentan', 'pioglitazone', 'flucloxacillin')
for case_name in selection:

    case_data = data[case_name]

    # Get SMARTCyp prediction
    response = requests.post('{0}/som_prediction'.format(MDSTUDIO_URL),
                             files={'ligand_file': case_data['mol2']}, data={'cyp': '3A4'})
    response_df = pandas.DataFrame.from_dict(response.json(), orient='index')

    # Get FAME3 prediction
    response = requests.post('{0}/predictFromFiles'.format(FAME_URL), files={'files': case_data['mol']},
                             data={'includeMDL': False}, headers={"accept" : "application/json"})
    response_df, famecutoff = process_fame_results(response.json(), response_df)

    # Get MetPred prediction
    response = requests.get('{0}/prediction'.format(METPRED_URL), params={'compound': case_data['mol']})
    response_df, metpredcutoff = process_metpred_results(response.json(), response_df)

    # Draw molecule
    response_df_prob = response_df[['Docking', 'SMARTCyp', 'FAME', 'MetPred', 'MetPred reaction']].fillna(0)
    rdmol = Chem.rdmolfiles.MolFromMol2File('{0}.mol2'.format(case_name), sanitize=False)
    dock_svg = show_predictions(rdmol, response_df_prob['Docking'])
    smart_svg = show_predictions(rdmol, response_df_prob['SMARTCyp'])
    fame_svg = show_predictions(rdmol, response_df_prob['FAME'], cutoff=famecutoff)
    metpred_svg = show_predictions(rdmol, response_df_prob['MetPred'], cutoff=metpredcutoff)

    # Display Pandas Dataframe and 2D depictions
    display(HTML('<h3 align="center">{0}</h3>'.format(case_name)))
    display(HTML(no_wrap_div.format(dock_svg, smart_svg, fame_svg, metpred_svg)))
    display(style_dataframe(response_df_prob, fame_cutoff=famecutoff))
```

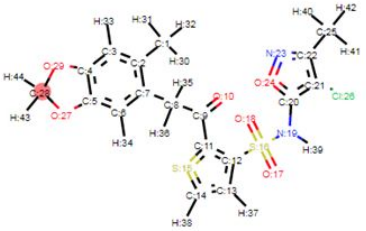
**sitaxentan**



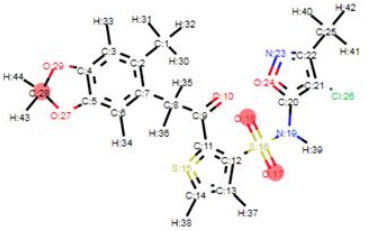
Docking



SMARTCyp



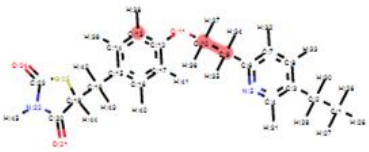
FAME 3



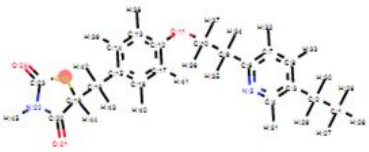
MetPred

| | Docking | SMARTCyp | FAME | MetPred | MetPred reaction |
|---|---|---|---|---|---|
| C.1 | 0.714 | 0.73 | 0.224 | 0 | 0 |
| C.11 | 0 | 0 | 0.0596 | 0 | 0 |
| C.12 | 0 | 0 | 0.052 | 0 | 0 |
| C.13 | 0.214 | 0.621 | 0.107 | 0 | 0 |
| C.14 | 0.214 | 0.699 | 0.176 | 0 | 0 |
| C.2 | 0 | 0 | 0.004 | 0 | 0 |
| C.20 | 0 | 0 | 0.056 | 0 | 0 |
| C.21 | 0 | 0 | 0.08 | 0 | 0 |
| C.22 | 0 | 0 | 0.184 | 0 | 0 |
| C.25 | 0 | 0.81 | 0.304 | 0 | 0 |
| C.28 | 0.214 | 1 | 0.892 | 0.179 | O-dealkylation; alkyl hydroxylation |
| C.3 | 1 | 0.6 | 0.0276 | 0 | 0 |
| C.4 | 0 | 0 | 0.004 | 0 | 0 |
| C.5 | 0 | 0 | 0.008 | 0 | 0 |
| C.6 | 1 | 0.6 | 0.02 | 0 | 0 |
| C.7 | 0 | 0 | 0.004 | 0 | 0 |
| C.8 | 0.786 | 0.73 | 0.148 | 0 | 0 |
| C.9 | 0 | 0 | 0.176 | 0 | 0 |
| CL.26 | 0.0714 | 0 | 0.048 | 0 | 0 |
| N.19 | 0.0714 | 0.514 | 0.036 | 0 | 0 |
| N.23 | 0 | 0.527 | 0.251 | 0 | 0 |
| O.10 | 0 | 0 | 0.0516 | 0 | 0 |
| O.17 | 0 | 0 | 0.0836 | 1 | oxidative desulfuration |
| O.18 | 0 | 0 | 0.0836 | 1 | oxidative desulfuration |
| O.24 | 0 | 0 | 0.152 | 0 | 0 |
| O.27 | 0 | 0 | 0.004 | 0 | 0 |
| O.29 | 0 | 0 | 0 | 0 | 0 |
| S.15 | 0.357 | 0.693 | 0.167 | 0 | 0 |
| S.16 | 0.643 | 0 | 0.052 | 0 | 0 |

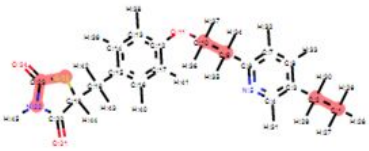
**pioglitazone**



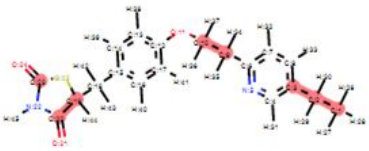
Docking



SMARTCyp



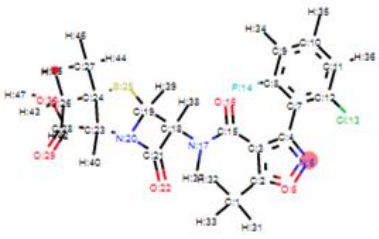
FAME 3



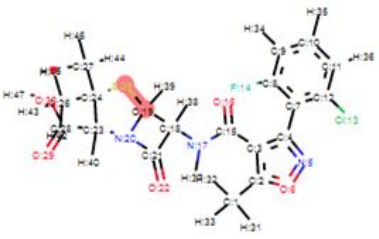
MetPred

| | Docking | SMARTCyp | FAME | MetPred | MetPred reaction |
|---|---|---|---|---|---|
| C.1 | 0 | 0.385 | 0.788 | 0.626 | carboxylation; alkyl hydroxylation |
| C.10 | 1 | 0.555 | 0.728 | 0.933 | alkyl hydroxylation; O-dealkylation |
| C.12 | 0 | 0 | 0 | 0 | 0 |
| C.13 | 0.833 | 0.447 | 0.008 | 0 | 0 |
| C.14 | 0.722 | 0.4 | 0 | 0 | 0 |
| C.15 | 0 | 0 | 0 | 0 | 0 |
| C.16 | 0.611 | 0.4 | 0 | 0 | 0 |
| C.17 | 0.556 | 0.447 | 0.008 | 0 | 0 |
| C.18 | 0.611 | 0.52 | 0.004 | 0 | 0 |
| C.19 | 0.667 | 0.598 | 0.204 | 0.616 | sulfide oxidation; alkyl hydroxylation |
| C.2 | 0 | 0.52 | 0.876 | 1 | alkyl hydroxylation |
| C.20 | 0 | 0 | 0 | 0.423 | alkyl hydroxylation; secondary amide hydrolysis (keep CO) |
| C.23 | 0 | 0 | 0.86 | 0.959 | alkyl hydroxylation; sulfide oxidation; amide hydrolysis (keep amine) |
| C.3 | 0 | 0 | 0 | 0.207 | aromatic hydroxylation |
| C.4 | 0.222 | 0.375 | 0.004 | 0 | 0 |
| C.6 | 0 | 0 | 0.012 | 0 | 0 |
| C.7 | 0.556 | 0.4 | 0 | 0 | 0 |
| C.8 | 0.389 | 0.4 | 0.008 | 0 | 0 |
| C.9 | 0.944 | 0.52 | 0.728 | 0.778 | alkyl hydroxylation |
| N.22 | 0 | 0.385 | 0.844 | 0 | 0 |
| N.5 | 0.333 | 0.456 | 0.032 | 0 | 0 |
| O.11 | 0 | 0 | 0.016 | 0 | 0 |
| O.21 | 0 | 0 | 0 | 0 | 0 |
| O.24 | 0 | 0 | 0 | 0 | 0 |
| S.25 | 0.5 | 1 | 0.996 | 0 | 0 |

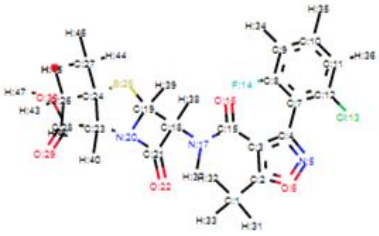
**flucloxacillin**



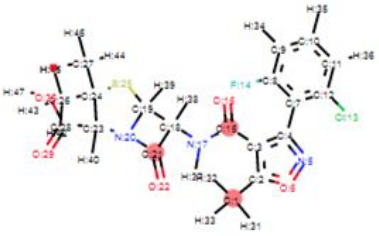
Docking



SMARTCyp



FAME 3



MetPred

| | Docking | SMARTCyp | FAME | MetPred | MetPred reaction |
|---|---|---|---|---|---|
| C.1 | 0.571 | 0.741 | 0.216 | 0.703 | alkyl hydroxylation |
| C.10 | 0 | 0.55 | 0.152 | 0 | 0 |
| C.11 | 0 | 0.528 | 0.04 | 0 | 0 |
| C.12 | 0 | 0 | 0 | 0 | 0 |
| C.15 | 0 | 0 | 0.112 | 0.346 | amide hydrolysis (keep amine); amine dehydrogenation |
| C.19 | 0.429 | 0.769 | 0.032 | 0 | 0 |
| C.2 | 0 | 0 | 0.06 | 0 | 0 |
| C.21 | 0 | 0 | 0.156 | 1 | amide hydrolysis (keep amine); tertiary amide hydrolysis (keep CO); alkyl hydroxylation |
| C.24 | 0 | 0 | 0.016 | 0 | 0 |
| C.26 | 0.571 | 0.496 | 0.04 | 0 | 0 |
| C.27 | 0.143 | 0.496 | 0.04 | 0 | 0 |
| C.28 | 0 | 0 | 0.012 | 0 | 0 |
| C.3 | 0 | 0 | 0.02 | 0 | 0 |
| C.4 | 0 | 0 | 0.0589 | 0 | 0 |
| C.7 | 0 | 0 | 0 | 0 | 0 |
| C.8 | 0.643 | 0 | 0.004 | 0 | 0 |
| C.9 | 0 | 0.528 | 0.04 | 0 | 0 |
| CA.18 | 0.286 | 0.695 | 0.008 | 0 | 0 |
| CA.23 | 0.286 | 0.669 | 0.012 | 0 | 0 |
| CL.13 | 0.357 | 0 | 0 | 0 | 0 |
| F.14 | 0.429 | 0 | 0 | 0 | 0 |
| N.17 | 0.714 | 0.496 | 0.016 | 0 | 0 |
| N.20 | 0.5 | 0.496 | 0.02 | 0 | 0 |
| N.5 | 1 | 0.482 | 0.196 | 0 | 0 |
| O.16 | 0 | 0 | 0.0276 | 0 | 0 |
| O.22 | 0 | 0 | 0.012 | 0 | 0 |
| O.29 | 0 | 0 | 0 | 0 | 0 |
| O.6 | 0 | 0 | 0.0636 | 0 | 0 |
| OXT.30 | 0 | 0 | 0.052 | 0 | 0 |
| S.25 | 0.286 | 1 | 0.048 | 0 | 0 |

OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

# Case Study

## Identification and Linking of Data related to AOPs of AOP-Wiki [**AOPLink**]

# SUMMARY

The Adverse Outcome Pathway (AOP) concept has been introduced to support risk assessment (Ankley et al., 2010). An AOP is initiated upon exposure to a stressor that causes a Molecular Initiating Event (MIE), followed by a series of Key Events (KEs) on increasing levels of biological organization. Eventually, the chain of KEs ends with the Adverse Outcome (AO), which describes the phenotypic outcome, disease, or the effect on the population.

In general, an AOP captures mechanistic knowledge of a sequence of toxicological responses after exposure to a stressor. While starting with molecular information, for example, the initial interaction of a chemical with a cell, the AOPs contain information of downstream responses of the tissue, organ, individual and population. Currently, AOPs are stored in the AOP-Wiki, a collaborative platform to exchange mechanistic toxicological knowledge as a part of the AOP-KB, an initiative by the OECD.

Normally, AOP development starts with a thorough literature search for existing knowledge, describing the sequence of KEs that form the AOP. However, the use of AOPs for regulatory purposes also requires detailed validation and linking to existing knowledge (Knapen et al., 2015; Burgdorf et al., 2017). Part of the development of AOPs is the search for data that supports the occurrence and biological plausibility of KEs and their relationships (KERs). This type of data can be found in literature, and increasingly in public databases.

The main goal of this case study is to establish the links between AOPs of the AOP-Wiki and experimental data to support a particular AOP. This will allow finding AOPs related to experimental data, and finding data related to a particular AOP.

# DESCRIPTION

## Implementation team

Coordination:

- Marvin Martens, Maastricht University, Department of Bioinformatics - BiGCaT
- Egon Willighagen, Maastricht University, Department of Bioinformatics - BiGCaT

Implementers:

- Risk assessors
- Modelers
- AOP developers
- Users of AOPs

## Case Study objective

The objective of this case study is to establish how existing AOPs on AOP-Wiki can be linked to experimental bioassay data. The approach here is to link assay data via assay types to key events (KEs) in the AOP.

For this case study we aim to develop:

- FAIR (Findable, Accessible, Interoperable and Reusable) version of AOP-Wiki;
- Identifier mappings for MIEs, KEs, and biological and chemical entities (genes, proteins, metabolites);
- Establish links between MIEs and KEs to biological assays and experimental data;
- Establish links between assays and biological and chemical entities;
- Establish interoperable databases.

## Risk assessment framework

The AOPLink case study covers a range of steps across different tiers of the SEURAT-1 risk assessment framework (Berggren et al., 2017). AOPLink allows finding relevant experimental data for given compounds and nanomaterials and KEs (Tier 0, step 3), identify biological processes affected by exposure to those chemicals supporting hypothesis generation (Tier 1, step 6), and using these sources of information to determine if an AOP can be applied to that chemical and if not what information is missing (Tier 3, step 9).

# Link to other case studies

With respect to the other OpenRiskNet case studies, AOPLink has a strong link to Datacure, as the primary goal of AOPLink is the search for experimental datasets related to AOPs of interest. Furthermore, AOPLink can take as input from **SysGroup** on similar chemicals (same group) in case no direct search results are found with the chemical of interest. Also, **TGX** may provide predicted data to complement experimental data, to support searching, and predicting the activation of a range of Molecular Initiating Events (MIEs). Because AOPLink may result in hypothesis and list KERs, these results can be passed to **ModelRX** for further prediction and read-across.

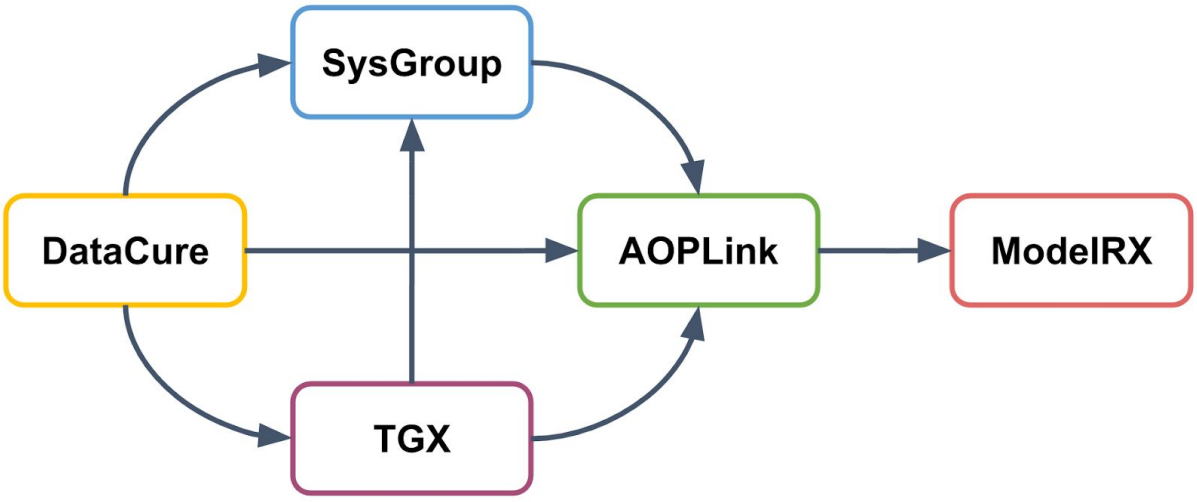



Figure 1: AOPLink and links to other case studies

# DEVELOPMENT

## Databases and tools

The following sets of repositories and services are used in the AOPLink case study.

- AOP-related repositories:
  - AOP-Wiki
  - AOP-DB
- Biological pathway databases/tools
  - WikiPathways
  - Reactome
- Experimental data repositories
  - diXa data warehouse
  - BioStudies
  - ArrayExpress
  - ToxCast
  - ToxRefDB
  - TG-GATEs
  - eNanoMapper
  - EPA Chemistry Dashboard
  - NORMAN Network
- Identifier mapping services
  - BridgeDb
  - ChemIdConverter
- Pathway analysis
  - PathVisioRPC

## Service integration

### Services integrated for AOPLink

AOP-Wiki:
The AOP-Wiki repository is part of the AOP Knowledge Base (AOP-KB), a joint effort of the US-Environmental Protection Agency and European Commission - Joint Research Centre. It is developed to facilitate collaborative AOP development, storage of AOPs, and therefore allow reusing toxicological knowledge for risk assessors. This Case Study has converted the AOP-Wiki XML data into an RDF schema, which has been exposed in a public SPARQL endpoint in the OpenRiskNet e-infrastructure.

EPA AOP Database (AOP-DB):
The EPA AOP-DB supports the discovery and development of putative and potential AOPs. Based on public annotations, it integrates AOPs with gene targets, chemicals, diseases, tissues, pathways, species orthology information, ontologies, and gene interactions. The AOP-DB facilitates the translation of AOP biological context, and associates assay, chemical and disease endpoints with AOPs (Pittman et al., 2018; Mortensen et al., 2018).

The AOP-DB won the first OpenRiskNet implementation challenge of the associated partner program and is therefore integrated into the OpenRiskNet e-infrastructure. After the conversion of the AOP-DB into an RDF schema, its data will be exposed in a Virtuoso SPARQL endpoint.

WikiPathways and Reactome:
WikiPathways is a community-driven molecular pathway database, supporting wide-spread topics and supported by many databases and integrative resources. It contains semantic annotations in its pathways for genes, proteins, metabolites, and interactions using a variety of reference databases, and WikiPathways is used to analyze and integrate experimental omics datasets (Slenter et al., 2017). Furthermore, human pathways from Reactome (Fabregat et al., 2018), another molecular pathway database, are integrated with WikiPathways and are therefore part of the WikiPathways RDF (Waagmeester et al., 2016). On the OpenRiskNet e-infrastructure, the WikiPathways RDF, which includes the Reactome pathways, is exposed via a Virtuoso SPARQL endpoint.

eNanoMapper:
The eNanoMapper database hosts nanomaterials characterization data and biological and toxicological information. It allows users to upload and explore data and information about nanomaterials through a REST web services API and a web browser interface, which is available in the OpenRiskNet e-infrastructure, using a newly developed Docker image.

BridgeDb:
In order to link databases and services that use particular identifiers for genes, proteins, and chemicals, the BridgeDb platform is integrated into the OpenRiskNet e-infrastructure. It allows for identifier mapping between various biological databases for data integration and interoperability (van Iersel et al., 2010).

PathVisioRPC:
To allow the analysis and visualization of transcriptomics or metabolomics data, PathVisioRPC (Bohler et al., 2015) will be used in AOPLink workflows. It is an XML-RPC interface, available for use in a variety of coding environments. It supports the use of pathways from WikiPathways for pathway statistics, exporting of results and providing data visualization on the pathways.

## Services provided by other case studies

### DataCure:

EdelweissData: Curated datasets are made available through the EdelweissData Explorer, the main data provisioning tool in the DataCure case study. It is a web-based data explorer tool that gives users the ability to filter, search and extract data through the use of API calls. The EdelweissData Explorer serves data from ToxCast, ToxRefDB, and TG-GATES.

ChemIdConverter: The ChemIdConverter allows users to submit and translate a variety of chemical descriptors, such as SMILES and InChI, through a REST API.

### SysGroup:

Grouping service that classifies a chemical or nanomaterial and provides structurally

and/or biologically similar chemical substances or compounds.

**TGX:**

API to access predicted data for the activation of a selective set of Molecular Initiating Events.

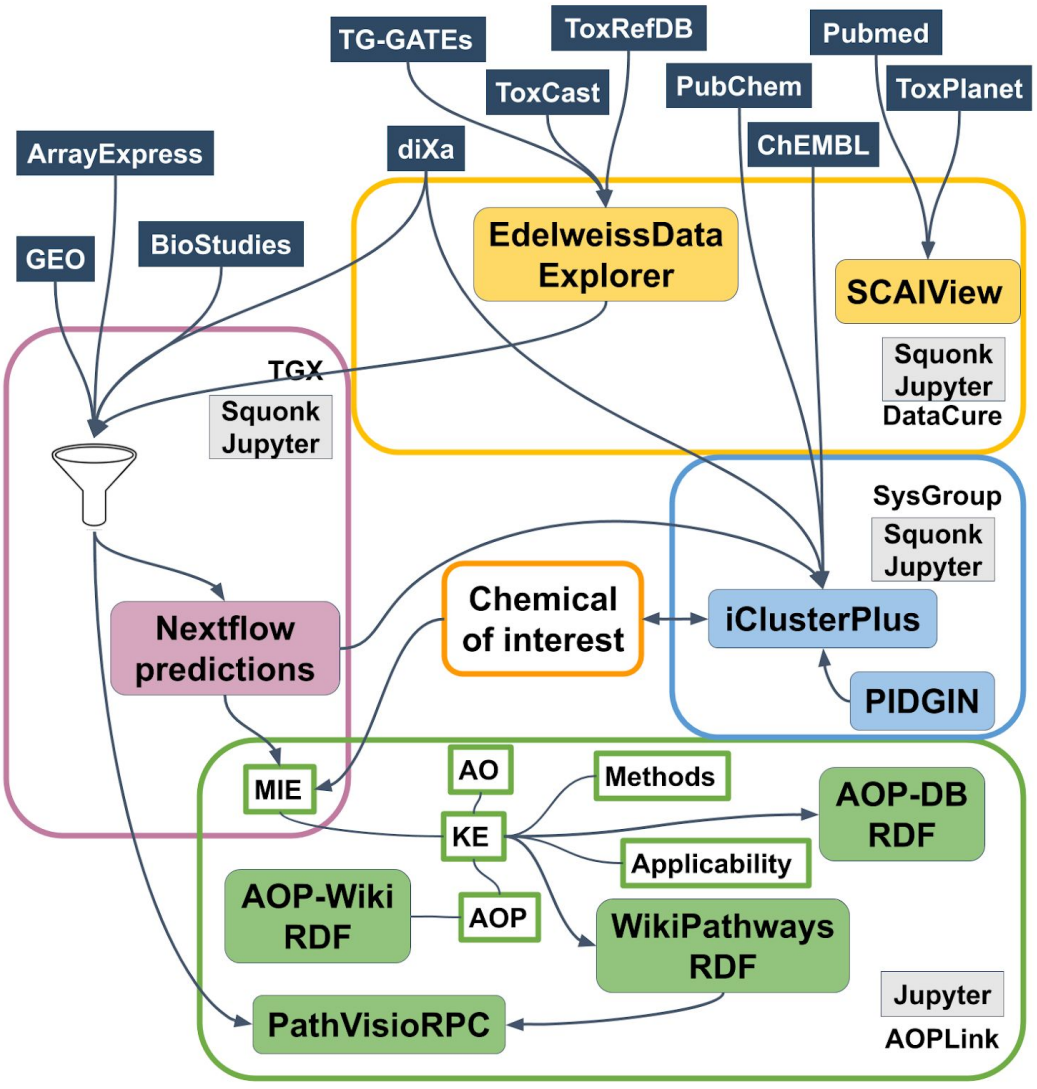



Figure 2: Network of services and databases from the case studies AOPLink, DataCure, SysGroup and TGX.

# Technical implementation

Jupyter Notebooks will be used to integrate the tools into reproducible workflows. Of the services, data from the AOP-Wiki, AOP-DB, and WikiPathways are available through SPARQL endpoints and are easily queried with the SPARQLwrapper python library. The other services, which have OpenAPI 2 or 3 definitions, are called through direct API calls. Combining the various tools and databases into workflows, the research goals are answered with modular, reproducible Jupyter Notebooks. For example, workflows are developed to find public experimental data that supports AOPs and to retrieve AOPs that are related to available data. Furthermore, we aim to develop complete data analysis workflows using WikiPathways and PathVisioRPC and relate the results to the knowledge captured in AOP-Wiki and AOP-DB.

# OUTCOMES

This case study focused mainly on the use of AOP knowledge, and extend it with additional information, or experimental data. The main repository in those analyses was the AOP-Wiki, which was converted into RDF, deployed in a Virtuoso SPARQL endpoint.

## AOPs linked to WikiPathways

One of the first analyses in the AOPLink case study was the assessment of the possible links between AOPs of AOP-Wiki, and the molecular pathways of WikiPathways. This task involved the identification of ontology usage for describing biological processes, and looking for the overlap of chemical coverage in both repositories. This exercise showed that few of the AOP-linked chemicals are found in WikiPathways, whereas 70% of all mapped genes are present in molecular pathways on WikiPathways. A manual assessment indicated that 67% of all low-level KEs, including molecular, cellular, tissue and organ KEs, can be linked (partially) to molecular pathways. [Martens M et al.]

## Workflow for finding data related to an AOP

One of the main questions to solve in AOPLink is the finding of data that supports an AOP of interest. To answer that, we have developed a Jupyter notebook that does that by using the AOP-Wiki RDF, AOP-DB RDF, BridgeDb, EdelweissData explorer, and WikiPathways services. The workflow, which was also presented during the workflow tutorial at the final workshop of OpenRiskNet, AOP 37 was selected as the AOP of interest.

First, the AOP-Wiki RDF was used to extract information about the AOP by using a variety of SPARQL queries that directly access the data through the SPARQL endpoint with the SPARQLWrapper library. Information of the AOP, such as the title, abstract, KEs and stressors, were extracted and written in a data frame, along with an AOP network that displays the connected AOPs.

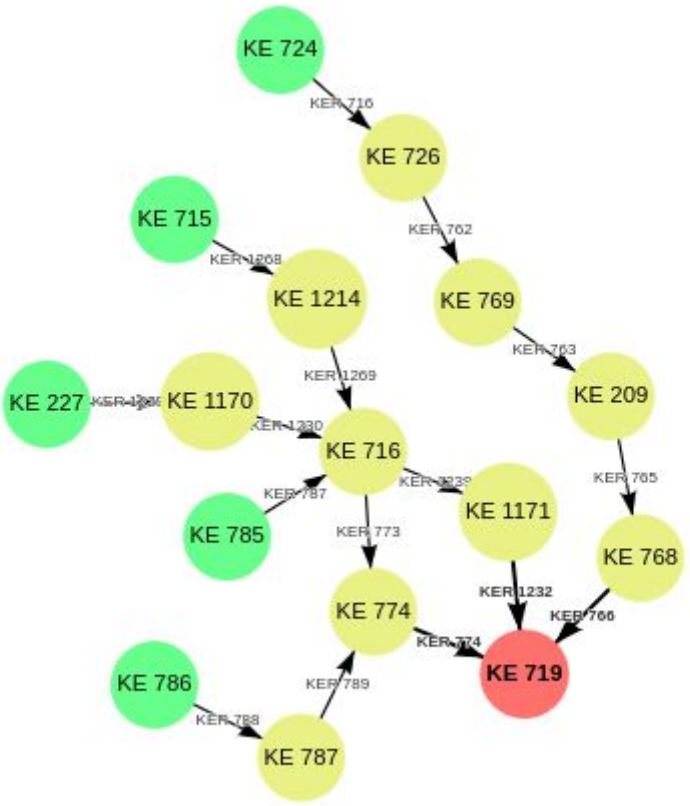



Figure 3: AOP network of AOP 37 with other AOPs extracted from AOP-Wiki RDF

```
sparqlquery = '''
PREFIX ncit: <http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#>
SELECT ?CAS_ID (fn:substring(?CompTox,33) as ?CompTox_ID)
WHERE{
?AOP_URI a aopo:AdverseOutcomePathway; ncit:C54571 ?Stressor.
?Stressor aopo:has_chemical_entity ?Chemical.
?Chemical cheminf:CHEMINF_000446 ?CAS_ID.
OPTIONAL {?Chemical cheminf:CHEMINF_000568 ?CompTox.}
FILTER (?AOP_URI = aop:'''+AOPid +''')}
'''
aopwikisparql.setQuery(sparqlquery)
aopwikisparql.setReturnFormat(JSON)
results = aopwikisparql.query().convert()

Chemical_names = {}

for result in results["results"]["bindings"]:
    try: Chemical_names[result["CAS_ID"]["value"]] =result["CompTox_ID"]["value"]
    except: pass

Chemdata = pd.DataFrame(columns=['CAS_ID','CompTox_ID'])
for CAS_ID in Chemical_names:
    Chemdata = Chemdata.append({
        'CAS_ID'        : CAS_ID,
        'CompTox_ID': Chemical_names[CAS_ID],
        }, ignore_index=True)
display(Chemdata)
```

| | CAS_ID | CompTox_ID |
|---|---|---|
| 0 | 117-81-7 | DTXSID5020607 |
| 1 | 25812-30-0 | DTXSID0020652 |
| 2 | 3771-19-5 | DTXSID8020911 |
| 3 | 41859-67-0 | DTXSID3029869 |
| 4 | 49562-28-9 | DTXSID2029874 |
| 5 | 50892-23-4 | DTXSID4020290 |
| 6 | 52214-84-3 | DTXSID8020331 |
| 7 | 637-07-0 | DTXSID3020336 |

Figure 4: Extracting chemicals from AOP-Wiki using SPARQL in a Jupyter notebook

Next, all chemical IDs were extracted from the found list of stressors, which were then used as an input for the ChemIdConverter to retrieve a variety of chemical descriptors, such as SMILES and InChI through the REST API.

```
compoundstable = pd.DataFrame(columns=['CAS_ID', 'Smiles', 'Image'])

# Fill "compounds" with the "smiles" by the compound name.
for compound in compounds:
    smiles = requests.get(chemidconvert + 'cas/to/smiles', params={'cas': compound}).json()['smiles']
    compoundstable = compoundstable.append({'CAS_ID': compound, 'Smiles': smiles, 'Image': smiles}, ignore_index=True)

def smiles_to_image_html(smiles):  # "smiles" shadows "smiles" from outer scope, use this function only in "to_html().
    """Gets for each smile the image, in HTML.
    :param smiles: Takes the "smiles" form "compounds".
    :return: The HTML code for the image of the given smiles.
    """
    return '<img style="width:250px" src="' + chemidconvert+'asSvg?smiles='+urllib.parse.quote(smiles)+'"/>'


# Return a HTML table of "compounds", after "compounds" is fill by "smiles_to_image_html".
HTML(compoundstable.to_html(escape=False, formatters=dict(Image=smiles_to_image_html)))
```

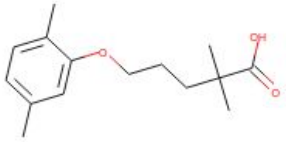
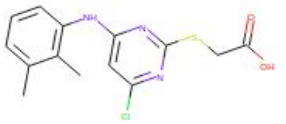
| | CAS_ID | Smiles | Image |
|---|---|---|---|
| 0 | 25812-30-0 | Cc1ccc(C)c(OCCCC(C)(C)C(O)=O)c1 |  |
| 1 | 50892-23-4 | Cc1cccc(Nc2cc(Cl)nc(SCC(O)=O)n2)c1C |  |

Figure 5: Chemical structures extracted through ChemIdConverter in a Jupyter notebook

To highlight that users of the AOP-Wiki are not required to learn SPARQL to access the data, the workflow shows how to access the content by using one of the predefined API calls, which was built using grlc.

```
import io
aopwiki_api = 'http://grlc.io/api/marvinm2/AOPWikiQueries/'
chemforaop = 'get-chemicals-for-aop'
headers={'accept': 'text/csv'}
result = requests.get(aopwiki_api+chemforaop+'?aopfilter='+AOPid, headers = headers)
resultClean = io.StringIO(result.content.decode("utf-8"))
df = pd.read_csv(resultClean, sep = ",")
df = df.drop(['callret-1'], axis=1)
display(df)
```

| | CASID |
|---|---|
| 0 | 117-81-7 |
| 1 | 25812-30-0 |
| 2 | 3771-19-5 |
| 3 | 41859-67-0 |
| 4 | 49562-28-9 |
| 5 | 50892-23-4 |
| 6 | 52214-84-3 |
| 7 | 637-07-0 |

Figure 6: Using the grlc API loaded with AOP-Wiki SPARQL queries to extract chemicals for AOP of interest in a Jupyter notebook

In order to extract all protein targets for the KEs of the AOP 37, the AOP-DB RDF was used. The data, which was converted to RDF and exposed in a SPARQL endpoint as part of the implementation challenge, was queried for all Entrez IDs linked to the AOP. Those were later used to extract all ToxCast Assay IDs from the AOP-DB RDF which have those genes as their target.

```
Assays = pd.DataFrame(columns=['Assay_ID', 'Assay_title', 'Entrez', 'Tissue', 'Species_name'])

for gene in Genes:
    sparqlquery = '''
    SELECT ?Assay_title ?Assay_ID ?Tissue ?Species_name WHERE{
    SELECT * WHERE{
    ?Assay a mmo:0000441; bao:BAO_0003064 ?Entrez_URI; rdfs:label ?Assay_title; foaf:group ?Assay
    SERVICE <http://aopwiki-rdf.prod.openrisknet.org/sparql/>{
    ?Species_URI dc:title ?Species_name.
    }
    FILTER (?Entrez_URI = ncbigene:'''+gene +''')}}
    '''
    aopdbsparql.setQuery(sparqlquery)
    aopdbsparql.setReturnFormat(JSON)
    results = aopdbsparql.query().convert()
    for result in results["results"]["bindings"]:
        Assays = Assays.append({'Assay_ID' :  result["Assay_ID"]["value"],
                                'Assay_title' :  result["Assay_title"]["value"],
                                'Tissue'      :  result["Tissue"]["value"],
                                'Species_name'     :  result["Species_name"]["value"],
                                'Entrez'        :  gene}, ignore_index=True)
display(Assays)
```

| | Assay_ID | Assay_title | Entrez | Tissue | Species_name |
|---|---|---|---|---|---|
| 0 | 269 | NVS_NR_hPPARa | 5465 | | Homo sapiens |
| 1 | 6 | ATG_TRANS | 5465 | liver | Homo sapiens |
| 2 | 6 | ATG_TRANS | 5465 | liver | Homo sapiens |

Figure 7: Using the AOP-DB to extract ToxCast assays related to our AOP of interest using SPARQL in a Jupyter notebook

Because Entrez IDs don't provide information about the gene name, or the species it belongs to, BridgeDb was used to map the Entrez IDs to HGNC and Ensembl IDs, showing that one of the four Entrez IDs was the human gene PPARA.

| | Entrez | HGNC | Ensembl |
|---|---|---|---|
| 0 | 5465 | PPARA | ENSG00000186951 |
| 1 | 403654 | | ENSCAFG00000000788 |
| 2 | 19013 | | ENSMUSG00000022383 |
| 3 | 25747 | | ENSRNOG00000021463 |

Figure 8: Results after Entrez identifier mapping for HGNC and Ensembl using BridgeDb

The next part focused on extracting transcriptomics data from TG-GATES using the EdelweissData explorer Python library, which accesses the EdelweissData API and queries for all datasets which were generated with the chemicals that were found earlier from the AOP-Wiki RDF. This resulted in a list of 181 datasets for different chemicals, species, dosings, and experimental design, of which we decided to focus on *in vivo* rat data with a high dose of exposure.

Finally, the data were analyzed by identifying the significantly affected molecular pathways. After querying all genes present in molecular pathways on WikiPathways for rats, pathway analysis was performed using the transcriptomics dataset from TG-GATES.

This indicated that 6 pathways from WikiPathways were significantly altered by the chemicals that activate the AOP 37.

| | Genes | nGenes | nSigGenes | percentSigGenes | Zscore |
|---|---|---|---|---|---|
| WP1286 | {108348148, 171341, 64305, 286954, 25279, 154985, 362228, 25428, 499422, 246245, 25429, 24294, 113992, 64352, 108348061, 24422, 116631, 681913, 140568, 364476, 64570, 494499, 246767, 288108, 29633, 316325, 100910462, 24404, 83783, 316435, 690050, 116686, 25086, 29326, 396551, 302302, 29328, 29725, 246248, 307859, 24426, 685402, 24192, 81924, 293779, 292915, 25256, 24421, 499302, 301264, 24297, 154516, 361510, 24902, 24861, 24424, 25315, 290623, 65030, 292155, 24298, 299566, 294449, 361631, 25426, 368066, 29680, 363618, 291770, 116632, 311257, 308445, 499689, 295934, 289197, 24267, 100910526, 65185, 303218, 574523, 58953, 25458, 58981, 103690051, 25147, 684979, 308511, 297029, 310848, 54246, 302489, 307838, 361718, 500257, 314694, 24912, 171072, 25427, 498489, 25355, ...} | 143 | 16 | 11.188811 | 5.152989 |
| WP1307 | {29367, 117243, 364975, 170670, 25363, 25288, 25330, 24849, 171155, 100911615, 94340, 311849, 113976, 29740, 361676, 113965, 25062, 289481, 298942, 50682, 79223, 114024, 25413, 24158, 117035, 64304, 311569, 25756, 24539, 140547, 25287, 117543, 24538, 25757, 25014} | 35 | 16 | 45.714286 | 5.152989 |
| WP372 | {29367, 117243, 364975, 170670, 25363, 25288, 25330, 24849, 171155, 100911615, 94340, 311849, 113976, 361676, 113965, 25062, 289481, 298942, 50682, 79223, 114024, 25413, 24158, 117035, 64304, 311569, 25756, 24539, 140547, 25287, 24538, 25757, 25014} | 33 | 14 | 42.424242 | 4.422721 |
| WP358 | {311743, 25317, 25054, 108348113, 25597, 81649, 170583, 58960, 24629, 24674, 497961, 60587, 303823, 311341, 140726, 66017, 309452, 170630, 365355, 299618, 691905, 171492, 170633, 170920, 314384, 170632, 24518, 306243, 500526, 116663, 296091, 79240, 116596, 170580, 297682, 315994, 24653, 363855, 170851, 286993, 81647, 311245, 314322, 25114, 25272, 24628, 293847, 84488, 257648, 287398, 361715, 29717, 116457, 81504, 25266, 140729, 308267, 29349, 50658, 24329, 58942, 24681, 25443, 24525, 24592, 29332, 501099, 63995, 362332, 304786, 300160, 25676, 100910732, 65179, 24577, 25388, 25625, 140926, 171337, 29568, 309361, 170538, 310533, 81737, 85384, 314436, 25233, 24493, 60661, 170579, 309295, 300980, 25636, 25112, 114856, 363875, 94202, 24559, 309165, 24680, ...} | 237 | 10 | 4.219409 | 2.962183 |
| WP2376 | {108348148, 29741, 81749, 171341, 81649, 363131, 58819, 24565, 58960, 301555, 24534, 170924, 57298, 54244, 83688, 689130, 365355, 24252, 85243, 81633, 300711, 83619, 64352, 305540, 117254, 108348061, 366960, 25319, 117263, 64191, 25283, 84006, 363855, 84509, 497932, 170851, 314322, 25260, 79130, 116686, 100361457, 314856, 289021, 50658, 29326, 24681, 29702, 24525, 24919, 300084, 85252, 64557, 24189, 24426, 103690044, 29568, 295549, 170538, 291084, 81822, 25256, 113894, 24421, 54349, 108348130, 29224, 681050, 301252, 65052, 29437, 497672, 24424, 24778, 25315, 81827, 29677, 24680, 170911, 24377, 366508, 367858, 25023, 313121, 116643, 289990, 25073, 24516, 499689, 289197, 24605, 24787, 60664, 116667, 293621, 361632, 24451, 84027, 25352, 246760, 25458, ...} | 167 | 8 | 4.790419 | 2.231914 |
| WP419 | {25541, 114024, 25287, 25413, 170670, 24158, 117035, 64304, 113976, 113956, 171142, 25363, 25288, 29740, 25757, 113965} | 16 | 8 | 50.000000 | 2.231914 |

Figure 9: Pathway analysis results using WikiPathways and TG-GATES data

```
WP1286  Metapathway biotransformation
WP1307  Fatty Acid Beta Oxidation
WP372   Beta Oxidation Meta Pathway
WP358   MAPK Signaling Pathway
WP2376  Nuclear factor, erythroid-derived 2, like 2 signaling pathway
WP419   Mitochondrial LC-Fatty Acid Beta-Oxidation
```

Figure 10: Pathway titles of significantly affected pathways from WikiPathways

**Further reading**:

Martens, M., Verbruggen, T., Nymark, P., Grafström, R., Burgoon, L. D., Aladjov, H., Torres Andón, F., Evelo, C.T., Willighagen, E. L. (2018). Introducing WikiPathways as a Data-Source to Support Adverse Outcome Pathways for Regulatory Risk Assessment of Chemicals and Nanomaterials. Frontiers in Genetics, 9, 661. doi:10.3389/fgene.2018.00661

# REFERENCES


1. Ankley, G. T., Bennett, R. S., Erickson, et al., (2010), Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment. Environmental Toxicology and Chemistry, 29: 730-741. doi:10.1002/etc.34
2. Dries Knapen, Lucia Vergauwen, Daniel L. Villeneuve, Gerald T. Ankley, The potential of AOP networks for reproductive and developmental toxicity assay development, Reproductive Toxicology, Volume 56, 2015, Pages 52-55, ISSN 0890-6238, https://doi.org/10.1016/j.reprotox.2015.04.003.
3. Tanja Burgdorf, Sebastian Dunst, Norman Ertych, Verena Fetz, Norman Violet, Silvia Vogl, Gilbert Schönfelder, Franziska Schwarz, and Michael Oelgeschläger, The AOP Concept: How Novel Technologies Can Support Development of Adverse Outcome Pathways, Applied In Vitro Toxicology 2017 3:3, 271-277. doi:10.1089/aivt.2017.0011.
4. Elisabet Berggren, Andrew White, Gladys Ouedraogo, et al.,, Ab initio chemical safety assessment: A workflow based on exposure considerations and non-animal methods, Computational Toxicology, Volume 4, 2017, Pages 31-44, ISSN 2468-1113, 10.1016/j.comtox.2017.10.001.
5. Pittman, M. E., Edwards, S. W., Ives, C., & Mortensen, H. M. (2018). AOP-DB: A database resource for the exploration of Adverse Outcome Pathways through integrated association networks. Toxicology and applied pharmacology, 343, 71–83. doi:10.1016/j.taap.2018.02.006
6. Mortensen, H.M., Chamberlin, J., Joubert, B. et al. Mamm Genome (2018) 29: 190. doi:10.1007/s00335-018-9738-7
7. Slenter DN, Kutmon M, Hanspers K, Riutta A, Windsor J, Nunes N, Mélius J, Cirillo E, Coort SL, Digles D, Ehrhart F, Giesbertz P, Kalafati M, Martens M, Miller R, Nishida K, Rieswijk L, Waagmeester A, Eijssen LMT, Evelo CT, Pico AR, Willighagen EL. WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research Nucleic Acids Research, (2017) doi:10.1093/nar/gkx1064
8. Fabregat, Antonio et al. “The Reactome Pathway Knowledgebase.” Nucleic acids research vol. 46,D1 (2018): D649-D655. doi:10.1093/nar/gkx1132
9. Waagmeester A, Kutmon M, Riutta A, Miller R, Willighagen EL, Evelo CT, et al. (2016) Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources. PLoS Comput Biol 12(6): e1004989. doi:10.1371/journal.pcbi.1004989
10. van Iersel, Martijn P et al. “The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services.” BMC bioinformatics vol. 11 5. 4 Jan. 2010, doi:10.1186/1471-2105-11-5
11. Bohler, A. *et al.,* “Automatically visualise and analyse data on pathways using PathVisioRPC from any programming environment.” BMC bioinformatics 16.1 (2015): 267. doi:10.1186/s12859-015-0708-8.

OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

# Case Study

## Toxicogenomics-based prediction and mechanism identification [**TGX**]

# SUMMARY

In this case study a transcriptomics-based hazard prediction model for identification of specific molecular initiating events (MIE) was foreseen based on (A) top-down and (B) bottom-up approaches.

The MIEs can include, but are not limited to: (1) Genotoxicity (p53 activation), (2) Oxidative stress (Nrf2 activation), (3) Endoplasmic Reticulum Stress (unfolded protein response), (4) Dioxin-like activity (AhR receptor activation), (5) HIF1 alpha activation and (6) Nuclear receptor activation (e.g. for endocrine disruption).

This case study focussed on two top-down approaches for genotoxicity prediction. The first approach resulted in the creation of a Nextflow-based workflow from the publication "A transcriptomics-based in vitro assay for predicting chemical genotoxicity in vivo" by Magkoufopoulou *et al.* (2012), thereby reproducing their work as proof of principle. The workflow for one of the prediction models described in the publication is shown in Figure 1.

The Nextflow-based workflow has been translated into a more generic approach, especially for step 1, forming the basis of the second top-down approach. In this approach transcriptomics data together with toxicological compound information were collected from multiple toxicogenomics studies and used for building a metadata genotoxicity prediction model.

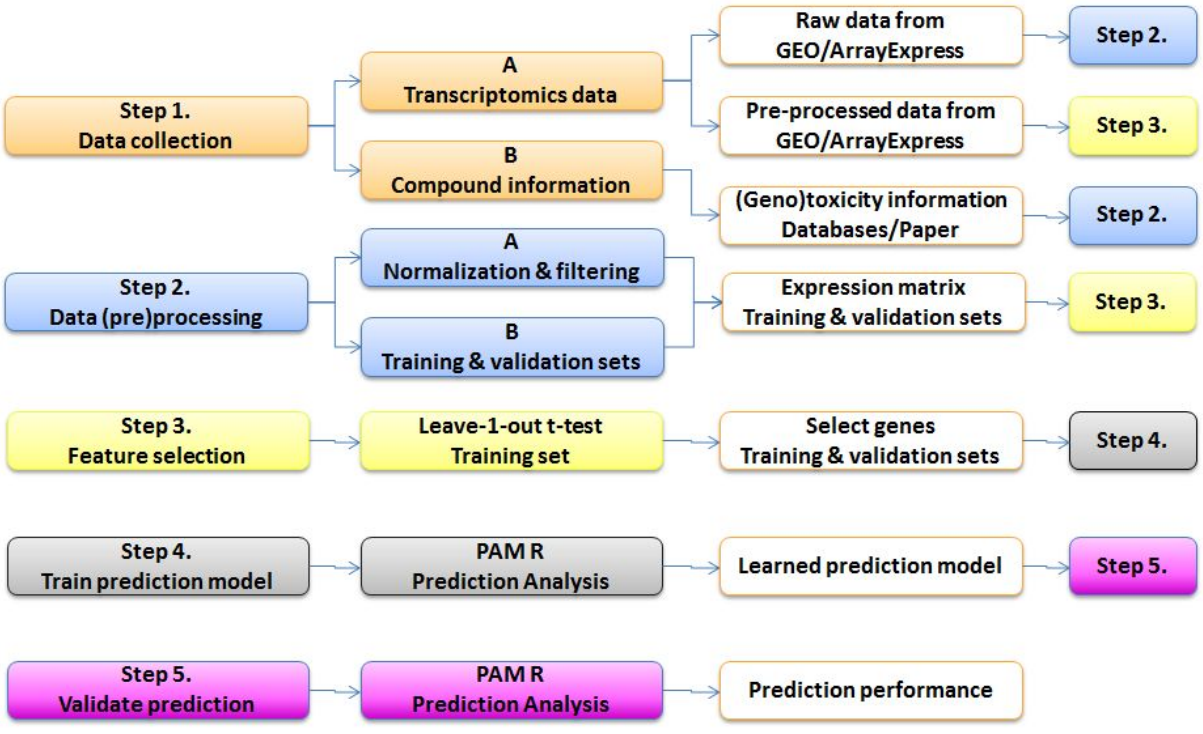



**Figure 1**. Workflow for genotoxicity prediction using whole genome transcriptomics data.

# DESCRIPTION

## Implementation team

Coordination:

- Danyel Jennen, Maastricht University, Department of Toxicogenomics

Other members:

- Jumamurat Bayjanov, Maastricht University, Department of Toxicogenomics
- Evan Floden, CRG

## Case Study objectives

- Creation of prediction models based on differentially regulated genes (top-down approach);
- Using knowledge of stress response pathways to integrate data sets for their activation or inhibition (bottom-up approach).

These two use cases are relevant for the top-down approaches:

- Reproducing the prediction models published by Magkoufopoulou et al 2012;
- Advanced predictions using as much data as possible from the diXa data warehouse[1] and other repositories giving free access to the data.

## Risk assessment framework

This case study is associated with all 3 tiers of the selected framework and in particular the following steps:

- Collection of support data;
- Identification of analogues / suitability assessment and existing data;
- Mode of Action hypothesis generation.

[1] http://wwwdev.ebi.ac.uk/fg/dixa/index.html

# DEVELOPMENT

## Databases and tools

**Databases**:

- diXa data warehouse (carcinoGENOMICs, Predict-IV), TG-GATEs, ArrayExpress/GEO, BioStudies.

**Tools**:

- Top-down: data normalisation tools, prediction tools such as Caret[2];
- Bottom-up: ToxPi.

## Technical implementation

Integration with other case studies is needed. TGX acquires information and data from the DataCure case study as well as through the services ToxPlanet and ToxicoDB of the Implementation Challenge winners Toxplanet and UHH, respectively. The results of TGX can feed into SysGroup, AOPLink and ModelRX.

Currently available services:

- Nextflow
  - Service to run Nextflow pipelines
  - Service type: Service, Workflow, Software
- Transcriptomics data from human, mouse, rat in vitro liver models
  - Repository for transcriptomics data from multiple in vitro human, rat and mouse toxicogenomics projects
  - Service type: Database / data source

[2] http://topepo.github.io/caret/index.html

# OUTCOMES

Outcome from this case study provides workflows for obtaining data which are suited for developing toxicity prediction models. This resulted in two top-down approaches for genotoxicity prediction.

## First top-down approach

A workflow from the earlier publication "A transcriptomics-based in vitro assay for predicting chemical genotoxicity in vivo" by Magkoufopoulou *et al.* (2012) [1] was developed, thereby reproducing their work as proof of principle. The workflow was created for one of the three approaches that were described in the study. There were some minor differences between the newly developed workflow and the original study, but overall the results of the original study were reproduced.

The workflow created using the Snakemake workflow manager is available from a GitLab software repository[3], where every step is clearly described in Snakefile to reproduce the approach described by Magkoufopoulou *et al.* (2012) [1] and was used as reference for the transfer into an OpenRiskNet-based solution. The repository also includes required scripts as well as description of the steps necessary to reproduce the results.

The Snakemake-based workflow was converted to a Nextflow-based workflow[4], named nf-toxomix, in order to make use of the harmonization and interoperability of OpenRiskNet. The Nextflow version uses containerised steps, thus making it easier to deploy on any cloud infrastructure, and applicable to OpenRiskNet Virtual Environments. Furthermore, the Nextflow-based workflow has been translated into a more generic approach so that it can be applied to other toxicogenomics studies.

## Second top-down approach

In this approach, transcriptomics data on human, mouse and rat *in vitro* liver cell models exposed to hundreds of compounds were collected from the diXa data warehouse, NCBI GEO and EBI's ArrayExpress using the workflow from the first top-down approach. The obtained datasets were merged per species. This was done manually because of differences in the description of the datasets, e.g. differences in used ontologies, different metadata file formats. For all the compounds used in the experiments genotoxic and carcinogenic information was gathered from literature and several databases, including ToxPlanet. After normalization of the transcriptomics data (per species) and gathering of the genotoxicity/carcinogenicity information, the data were ready to be fed into the prediction models of ModelRX.

The transcriptomics data from the human, mouse, rat *in vitro* liver cell models and the toxicological information are available through OpenRiskNet[5].

[3] https://gitlab.com/bayjan/openrisknet_magkoufopoulou
[4] https://github.com/openrisknet/nf-toxomix
[5] https://gitlab.com/bayjan/openrisknet_meta_analysis_data

# Related resources

**Use of Nextflow tool for toxicogenomics-based prediction and mechanism identification in OpenRiskNet e-infrastructure**

Evan Floden

27 May 2019 | Webinar

**OpenRiskNet Part II: Predictive Toxicology based on Adverse Outcome Pathways and Biological Pathway Analysis**

Marvin Martens, Thomas Exner, Nofisat Oki, Danyel Jennen, Jumamurat Bayjanov, Chris Evelo, Tim Dudgeon, Egon Willighagen

28 August 2019 | Poster

**Meta-analysis for genotoxicity prediction using data from multiple human in vitro cell models**

Jumamurat R. Bayjanov Jos Kleinjans Danyel Jennen

12 Sep 2018 | Poster

**Big Data in Toxicogenomics: Towards FAIR predictions**

Danyel Jennen

26 Jul 2018 | Presentation ICCA 2018

# REFERENCES

1. Magkoufopoulou C, Claessen SM, Tsamou M, Jennen DG, Kleinjans JC, van Delft JH. A transcriptomics-based in vitro assay for predicting chemical genotoxicity in vivo. Carcinogenesis. 2012 Jul;33(7):1421-9. doi:10.1093/carcin/bgs182.

OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

# Case Study

## Reverse dosimetry and PBPK prediction [**RevK**]

# SUMMARY

This case-study demonstrates and documents the use of a web interface to physiologically-based pharmacokinetic models for forward and reverse dosimetry calculations. Forward calculations compute internal concentrations from given exposure doses. Reverse calculations compute exposure doses from internal concentrations or measured biomarker levels (*e.g.*, urine concentration data). The result of those calculations can be used in risk assessments to help with *in vitro* to *in vivo* extrapolations or interspecies extrapolations.

Three tools have been developed for this case-study at NTUA and have been integrated into the OpenRiskNet infrastructure through the Jaqpot web-based computational platform. More specifically, the popular high-throughput toxicokinetic (*httk)* R package and the *PKSim* software tool for whole-body physiologically based pharmacokinetic modeling were integrated, but we also developed infrastructure for developing and deploying user-defined model.

For each of these three web tools, simulations are performed and results are presented for reference chemicals or drugs, namely Imazalil for the *httk* model, Diazepam and Chlorpyrifos for showcasing the In-house R PBPK workflow and Theophylline for the *PKSim* model. The exposure scenarios chosen are in the range of corresponding environmental or therapeutic levels.

Finally, a brief overview describing how to access custom-made PBPK models and run simulations through the Jaqpot Graphical User Interface (GUI) is presented in the Annex.

# DESCRIPTION

## Implementation team

| CS leader | Team |
|---|---|
| Frederic Bois / Celine Brochot (INERIS) | Haralambos Sarimveis, Periklis Tsiros, Pantelis Karatzas, Philip Doganis (NTUA)<br><br>Cleo Tebby (INERIS) |

- The code and web implementation was developed by NTUA
- The case-studies' simulations were run by INERIS and NTUA
- Code and technical service documentation was provided by NTUA. INERIS documented the users' operations and results of the case-study demonstrations.

## Case Study objective

The objective of this case study is to demonstrate and document the capabilities of the OpenRiskNet-developed web-services for Physiologically Based PharmacoKinetic (PBPK) modeling with illustration of both forward and reverse dosimetry predictions.
PBPK models offer a methodology for predicting the internal distribution and exposure of a compound in an organism. Their nature is mechanistic; they consist of compartments representing real organs and tissues, whose number varies based on the target substance, species, administration route and available information. A common approach is to incorporate in the model the main body tissues, i.e. brain, heart, kidney, skin, spleen, liver, lung, gut, bone, adipose and muscle (Jones et al., 2013). Nevertheless, the dimensionality of a PBPK model can be reduced using lumping methods (Pilari et al. 2010, Nestorov et al., 1998) . In most cases, PBPK models are utilised for describing the kinetics of a substance in the whole body of a species, thus such models are more formally called "whole body physiologically-based pharmacokinetic" (WBPBPK) models. However, there are models developed to describe in detail the kinetics of a specific organ or body area, which is divided into separate subcompartments. This modeling approach is called "partial" PBPK models (Sturm, 2007).
PBPK models have inherent advantages due to their mechanistic nature. Firstly, they enable predictions of concentration/mass profiles of individual organs and not just plasma. In addition, their relation with physiology and modularity facilitate the integration of literature information, making predictions prior to *in vivo* experiments possible (Nestorov, 2003). Lastly, their biggest advantage is the ability to perform inter-species (e.g. from rat to human) or intra-species (e.g. from adults to children) extrapolation through scaling methods.

## Risk assessment framework

The application frameworks are, for example: REACH risk assessments; SEVESO II directive on safety around industrial plants; Internal chemical, cosmetic, or pharmaceutical company assessments of workers' safety, or consumer's safety. All those require integration and extrapolation of *in vitro* and/or *in vivo* data on animals to assess human risks.

# DEVELOPMENT

## Databases and tools

We use open source software able to implement PBPK models within the Jaqpot platform (Chomenidis et al., 2017): *httk* package (Pearce et al., 2017), *PKSim* (Willmann et al., 2003), and an in-house R client for custom PBPK modelling. The *Jaqpot* biokinetic services are used to publish the PBPK models as web services. Service clients are developed in the R language. Databases of parameter values are provided by the *httk R* package, and the *PKSim* model.

## Technical implementation

**Implementation of the chosen PBPK model as web services**:

PBPK models for a specific class of chemicals and animal species can be selected by the user from a particular PBPK modelling environment (*e.g.*, *httk* in *R*, *PKSim*).

The chosen PBPK model is exposed as a web service using the Jaqpot modelling platform. This is possible through the Jaqpot Protocol of Data Interchange (JPDI) which allows to dynamically and seamlessly incorporate practically any algorithmic implementation into Jaqpot. The protocol specifies the form of data exchange between Jaqpot services and third party algorithm web service implementations. The Jaqpot framework already provides wrappers for the R language and the Python language. Integration with R is made possible through the OpenCPU system, which defines an HTTP API for embedded scientific computing based on R, although this approach could easily be generalized to other computational back-ends (Ooms, 2014). OpenCPU acts as a wrapper to R that is readily able to expose R functions as RESTful HTTP resources. The OpenCPU server takes advantage of multi-processing in the Apache2 web server to handle concurrency. This implementation uses forks of the R process to serve concurrent requests immediately with little performance overhead. By doing so it enables access to those functions on simple HTTP calls converting R from a standalone application to a web service.

**Demonstration of PBPK models that have been exposed as web services**:

The three simulation tools (*httk*, *PKSim* and user-specified) are demonstrated with Imazalil, Theophylline, Diazepam and Chlorpyrifos in rainbow trout respectively.

For Imazalil and Theophylline, we start by identifying relevant human exposures (*e.g.* from ExpoCast, or published literature) to be used in forward dosimetry. For Diazepam and Chlorpyrifos, reverse dosimetry is examined; we identify (*e.g.* from the US NHANES database, or the scientific literature) typical blood or urine concentrations found in humans to be used as input to the exposure dose reconstruction.

Each model is parameterized using user-specified or pre-programmed tabulated physiological data. For forward dosimetry predictions, each model is run with the given exposure scenario to predict internal concentrations after 24 hours, while for reverse dosimetry, the model is run forward iteratively with user set exposures so as to match the input biomarkers (that is: manually invert the model). The external exposure level leading to data-matching biomarker level is recorded as final estimate.

# OUTCOMES

In this section, several implementations of this case study are described:

The first implementation uses *httk* and Imazalil. We describe all the steps required to develop the models as web services through the Jaqpot API or the Jaqpot GUI.

The second is a generic OpenRiskNet framework , which can be used with custom-made PBPK models. Two examples are provided, a PBPK model for Diazepam in humans, and a generic (i.e. not substance-specific) PBPK model in fish. In the case of diazepam, the tools were used to analyse biomonitoring data regarding diazepam blood levels in drivers. In the case of the fish PBTK model, exposure levels which lead to in vitro effects on biomarkers in liver were estimated.

The last implementation is the integration of a PBPK model for Theophylline, originally developed in the PKSim software. We describe all the steps required to develop the model as a web service through the Jaqpot API.

We are providing all the steps required to perform dosimetry through the Jaqpot GUI using the custom-made model as examples.

The results of this case-study demonstrate that the OpenRiskNet framework can be used as a central e-platform for the biokinetics community, where the users can publish, share, search and use PBPK models.

# Integration of *httk* in OpenRiskNet Infrastructure through the Jaqpot modelling platform

High-Throughput toxicokinetics *(httk*) is an R package for simulation and analysis of chemical toxicokinetics. httk has been integrated in the Jaqpot platform and is accessible from the OpenRiskNet infrastructure. The goal of this tutorial is to demonstrate how to obtain toxicokinetics predictions from the httk package using Jaqpot services and functionalities through both API and GUI.

The following httk parameters are supported:

1. “chem.name”, which is the chemical name of the compound under study
2. “species”, which is the species of interest (either "Rat", "Rabbit", "Dog", "Mouse", or default "Human")
3. “days”, the length of the simulation in days,
4. “dose”, the amount of the single dose specified (mg/kg/day) with the default value being NULL.

We use the example of administering Imazalil to a human (BW = 70 kg) also initializing other basic parameters based on the multiple compartment model introduced in Kilford et al. (2008). The compartments used in this model are the gutlumen, gut, liver, kidneys, veins, arteries, lungs, and the rest of the body. The extra compartments include the amounts or concentrations metabolized by the liver and excreted by the kidneys through the tubules.

## Accessing *httk* through Jaqpot API

The following steps should be followed using the Swagger JaqPot API at https://api-jaqpot.prod.openrisknet.org/jaqpot/swagger/:

1) Produce a JaqPot resource that implements the httk model as a web service

a. Use the **POST/biokinetics/httk/createmodel** method.

Insert the following information in the parameter fields:

Title and description: Any title and description is fine

Parameter string:
{"chem.name":["imazalil"],"species":["Human"],"days":[10],"dose":[10]}

Press the ‘Try it out!’ button. In the field response body a json string will appear. There the user can find the task ID.

b. Use the **GET /task/{id}** method. Copy the task ID in the relevant field and press the ‘Try it out!’ button. If the status at the end of the

response body is not COMPLETED wait for a few seconds and try again. The model has been created and its complete URI is shown in the response body. For the next step store the model id. This is the id appearing after model/ in “result” in the response body.

The raw model together with the initial parameters of the multiple compartment model are returned.

2) Use the model for obtaining drug concentration-time profiles

a. Use the **POST biokinetics/httk/model/{id}** method in the biokinetics section. Submit the model id obtained before and then press the button ‘Try it out!’. A task ID is created.

b. Use the **GET /task/{id}** method. Copy the task ID in the relevant field and press the ‘Try it out!’ button. The dataset with drug concentration –time data has been created. For the next step store the dataset id. This is the id found after dataset/ in the “result” section of the response body.

c. Use the **GET /dataset/{id}** method: copy the dataset id in the relevant field and press ‘Try it out!’. In the Response Body, drug concentration values are shown along with corresponding time points (in days), and AUC (area under the curve) of the plasma concentration values.

These are the first lines of the result giving concentrations in uM:

"Substance","Agutlumen","Atubules","Cplasma","Cart","Ckidney","Crest","Cliver","Cgut","Ametabolized","Cven","time","Clung","AUC"

"1","2355.4749","0","0","0","0","0","0","0","0","0","0","0","0"

"2","1365.8025","0.1105","13.1333","9.2261","59.1467","2.8637","277.1605","258.6781","0","9.5873","0.0104","9.3056","0.0598"

"3","791.9496","0.5618","20.3668","14.7784","106.6832","11.1376","356.6631","205.9325","0","14.8678","0.0208","14.501","0.2427"

"4","459.2054","1.1426","21.3648","15.6043","115.3854","19.338","318.4675","156.002","0","15.5963","0.0312","15.2265","0.4629"

"5","266.2664","1.7291","20.5252","15.0134","111.6516","25.5365","259.9901","120.0944","0","14.9834","0.0417","14.6313","0.6817"

"6","154.3925","2.2885","19.4394","14.2189","105.7722","29.7367","210.7735","96.3627","0","14.1908","0.0521","13.8573","0.8898"

"7","89.5232","2.8195","18.5569","13.5674","100.7951","32.4253","175.7869","81.3744","0","13.5466","0.0625","13.2273","1.0874"

"8","51.9093","3.3285","17.938","13.1087","97.2449","34.0891","152.7805","72.164","0","13.0947","0.0729","12.7851","1.2773"

"9","30.0992","3.8225","17.5331","12.8081","94.9023","35.0962","138.2902","66.602","0","12.7992","0.0833","12.4959","1.4619"

"10","17.4528","4.3068","17.2782","12.6186","93.42","35.6962","129.3984","63.2808","0","12.6131","0.0938","12.3137","1.6431"

Other examples:

{"chem.name":["bisphenol A"],"species":["Rat"],"days":[10],"dose":[10]}

{"chem.name":["imazalil"],"species":["Rat"],"days":[10],"dose":[10]}

{"chem.name":["Acetochlor"],"species":["Rat"],"days":[10],"dose":[10]}

## Acessing *httk* through Jaqpot Graphical User Interface

A user interface for the specific httk services has been developed and is integrated into the overall Jaqpot GUI (https://ui-jaqpot.prod.openrisknet.org/). After login, the user can use the navigation bar of Jaqpot UI and select the httk application (Figure 1).

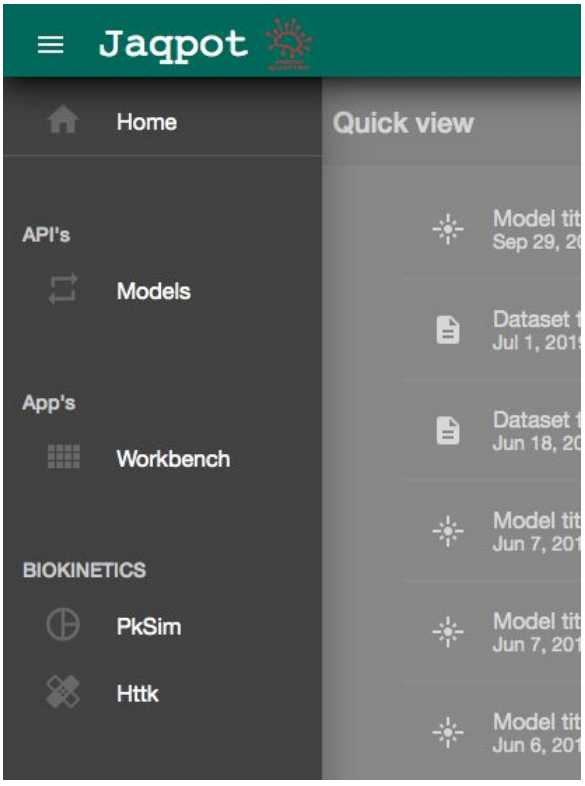



**Figure 1.** Jaqpot's Navigation bar.

The user is subsequently directed to the landing page of the *httk* Jaqpot web service. First the user is requested to complete the simulation information, which comprises 7 steps (Figure 2). Specifically, these steps are:

1. Select species; human and rat compose the available choices.
2. Select the chemical. This is done through an autocomplete system. After pressing a letter, all available chemicals that start with this letter are loaded. The user can continue the letter filling process until the available choices are narrowed down to the compound of interest.
3. Set the dose administered
4. Set the duration of the simulation
5. Provide a title for the simulation for archiving purposes
6. Provide a short description for the simulation
7. Create the model and obtain the predictions

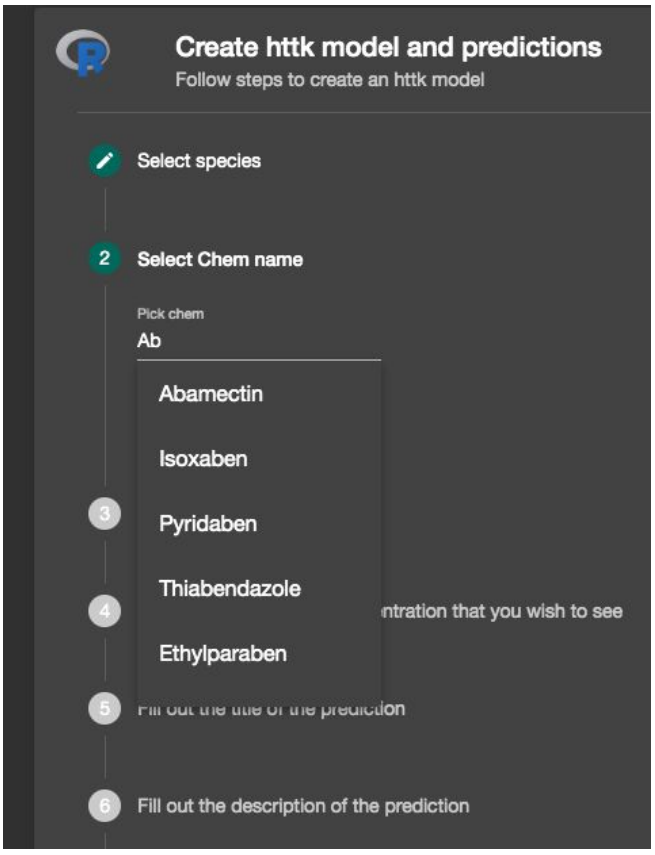
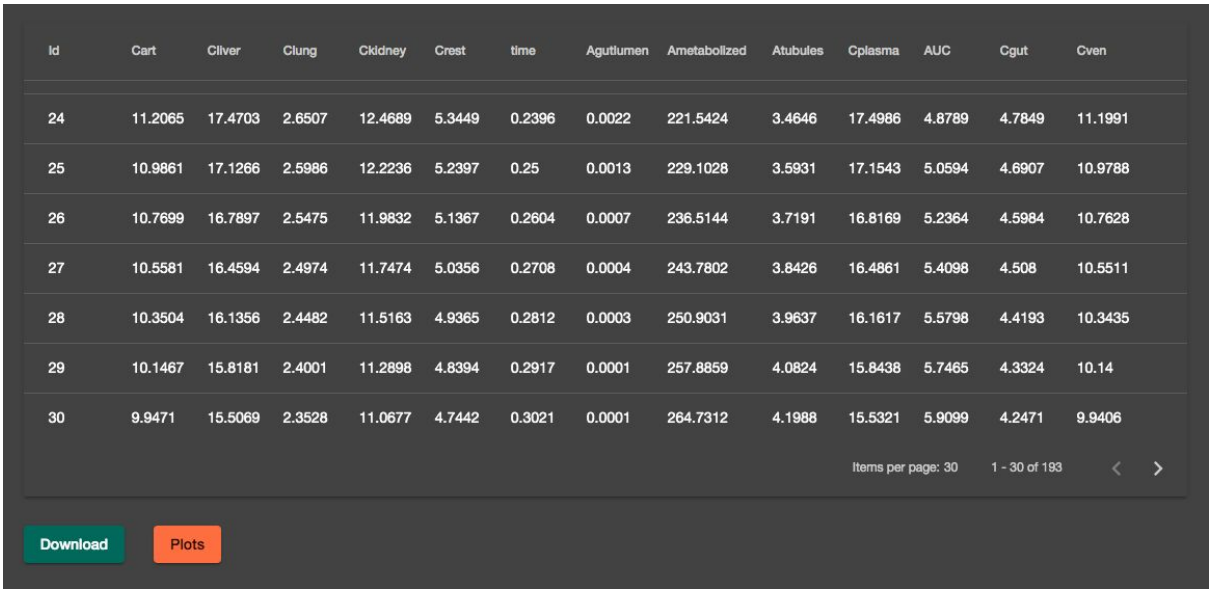



**Figure 2.** Steps of the httk model creation process.



| Id | Cart | Cliver | Clung | Ckidney | Crest | time | Agutlumen | Ametabolized | Atubules | Cplasma | AUC | Cgut | Cven |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 24 | 11.2065 | 17.4703 | 2.6507 | 12.4689 | 5.3449 | 0.2396 | 0.0022 | 221.5424 | 3.4646 | 17.4986 | 4.8789 | 4.7849 | 11.1991 |
| 25 | 10.9861 | 17.1266 | 2.5986 | 12.2236 | 5.2397 | 0.25 | 0.0013 | 229.1028 | 3.5931 | 17.1543 | 5.0594 | 4.6907 | 10.9788 |
| 26 | 10.7699 | 16.7897 | 2.5475 | 11.9832 | 5.1367 | 0.2604 | 0.0007 | 236.5144 | 3.7191 | 16.8169 | 5.2364 | 4.5984 | 10.7628 |
| 27 | 10.5581 | 16.4594 | 2.4974 | 11.7474 | 5.0356 | 0.2708 | 0.0004 | 243.7802 | 3.8426 | 16.4861 | 5.4098 | 4.508 | 10.5511 |
| 28 | 10.3504 | 16.1356 | 2.4482 | 11.5163 | 4.9365 | 0.2812 | 0.0003 | 250.9031 | 3.9637 | 16.1617 | 5.5798 | 4.4193 | 10.3435 |
| 29 | 10.1467 | 15.8181 | 2.4001 | 11.2898 | 4.8394 | 0.2917 | 0.0001 | 257.8859 | 4.0824 | 15.8438 | 5.7465 | 4.3324 | 10.14 |
| 30 | 9.9471 | 15.5069 | 2.3528 | 11.0677 | 4.7442 | 0.3021 | 0.0001 | 264.7312 | 4.1988 | 15.5321 | 5.9099 | 4.2471 | 9.9406 |

Items per page: 30 1 - 30 of 193

Download Plots

**Figure 3.** Tabular format of model predictions.

The results are presented in a tabular format (Figure 3). For a better understanding of the process kinetics, the user is given the ability to visualise the results by plotting the concentration-time profiles. This is realised by pressing the “Plots” button, which opens a window that requests the user to set the x and y-axis by selecting the compartments of interest through a drag-and-drop gesture (Figure 4). Figure 5 presents the plots resulting after selecting the compartments of Figure 4, namely the concentration of the gut, liver and rest-of-the-body compartments.

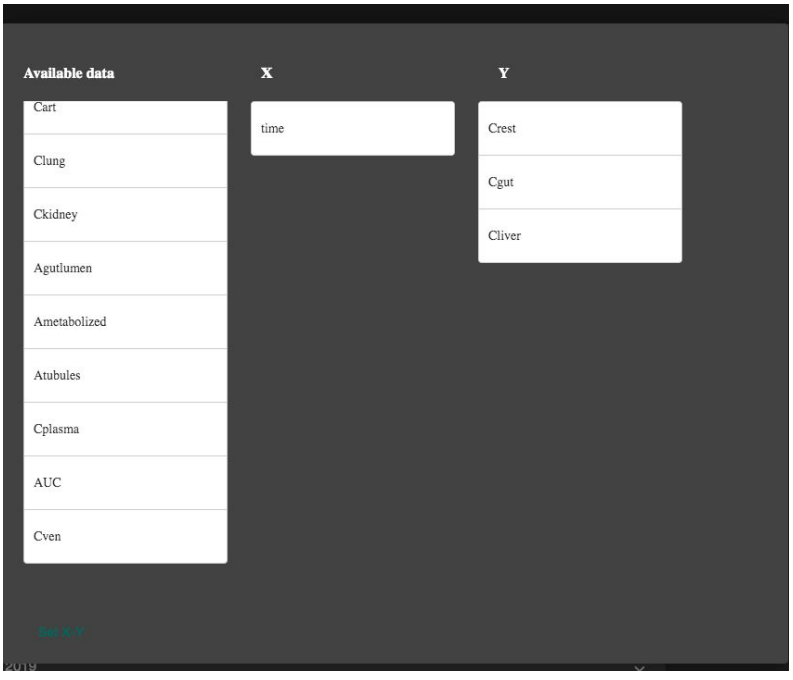



**Figure 4.** Pop-up window for setting up plots.

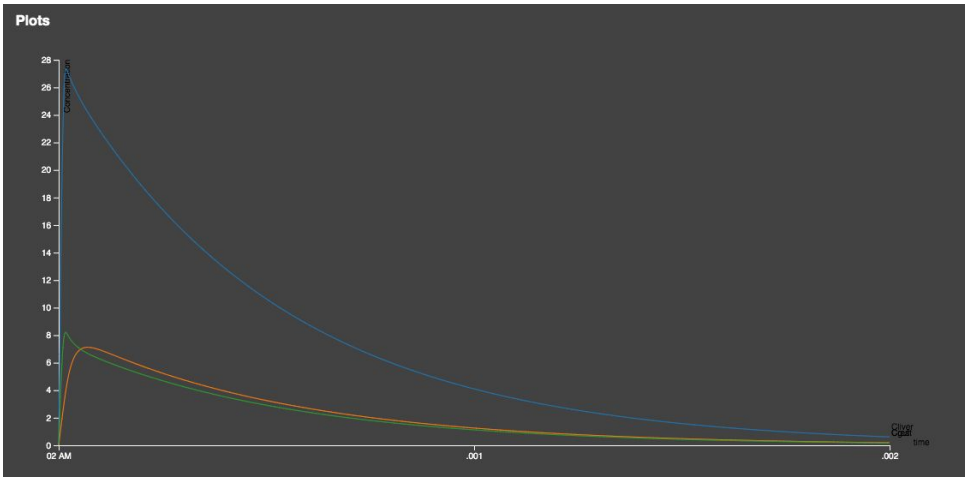
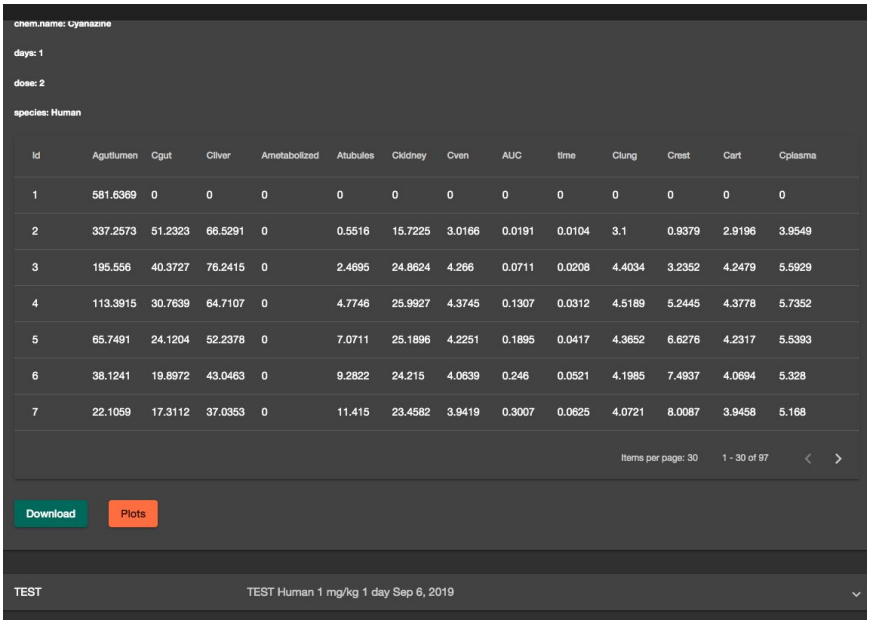



**Figure 5.** Plots generated in the httk web service GUI.

All information generated during the in-silico experiment is archived for accessing it at a later point in time, so that recreation of the same model, and thus needless spent of computational resources, is avoided. If a user wishes to revisit archived model predictions, pressing the “Previous predictions” makes all previous predictions and models available.




| Id | Agutlumen | Cgut | Cliver | Ametabolized | Atubules | Ckidney | Cven | AUC | time | Clung | Crest | Cart | Cplasma |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 581.6369 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 337.2573 | 51.2323 | 66.5291 | 0 | 0.5516 | 15.7225 | 3.0166 | 0.0191 | 0.0104 | 3.1 | 0.9379 | 2.9196 | 3.9549 |
| 3 | 195.556 | 40.3727 | 76.2415 | 0 | 2.4695 | 24.8624 | 4.266 | 0.0711 | 0.0208 | 4.4034 | 3.2352 | 4.2479 | 5.5929 |
| 4 | 113.3915 | 30.7639 | 64.7107 | 0 | 4.7746 | 25.9927 | 4.3745 | 0.1307 | 0.0312 | 4.5189 | 5.2445 | 4.3778 | 5.7352 |
| 5 | 65.7491 | 24.1204 | 52.2378 | 0 | 7.0711 | 25.1896 | 4.2251 | 0.1895 | 0.0417 | 4.3652 | 6.6276 | 4.2317 | 5.5393 |
| 6 | 38.1241 | 19.8972 | 43.0463 | 0 | 9.2822 | 24.215 | 4.0639 | 0.246 | 0.0521 | 4.1985 | 7.4937 | 4.0694 | 5.328 |
| 7 | 22.1059 | 17.3112 | 37.0353 | 0 | 11.415 | 23.4582 | 3.9419 | 0.3007 | 0.0625 | 4.0721 | 8.0087 | 3.9458 | 5.168 |

**Figure 6.** “Archive/history” page

# Integration of custom made PBPK models in the Jaqpot modelling platform

Besides the integration of httk and PKSim, we offer the option to the users to create, upload and share a custom-made PBPK model through the Jaqpot infrastructure. A user-friendly R client has been developed that allows model creators to expose their PBPK models as ready-to-use web services on the Jaqpot platform. The R client utilizes the deSolve package (R language) for solving differential equations. The deployment process is realised through a function with which the user needs to specify a series of components and it has been designed in a way that offers extended flexibility on the models to be uploaded (Figure 7). Specifically, the first component, 'user.input', should be a list containing the names of the input features, which the end-user needs to fill in. 'predicted.feats' is again a list consisting of the names of the predicted features that will be the final output on the Jaqpot GUI. The next element, 'create.params', is a function that receives as input the input of the end-user and transforms it according to the needs of the model. Following that, 'create.inits' and 'create.events' receive as input the output of create.params and create the initial conditions of the ODEs and a dataframe containing events that force changes on the state variables of the ODE system, respectively. The ability to use a custom function inside the ode function ('ode.func') is provided by 'custom.func', while 'ode.func' is a function that is forwarded to the 'ode' function of the R package 'deSolve' and contains the ODEs. The function is requested in the following format: ode.func(integration.time, initial.values, parameters, custom.func). Finally, the user can select a specific solver from the ones provided in 'deSolve' package through the 'method' argument and can also pass additional arguments (e.g. 'rtol' for changing the relative tolerance of the solution) down to the solver through the three dots R argument ('...').

```
> deploy.pbpk(user.input, predicted.feats, create.params, create.inits, create.events,
+                         custom.func, PBPK.model, method = "lsodes")
Base path of jaqpot *e.g.: https://api.jaqpot.org/ : https://api.jaqpot.org/
Please choose authentication method ([1]=login / [2]=Provide Api Key): 1
Username: PeriklisTs
Title of the model: Toy Example()
Short description of the model:This is a test model
[1] "Model created. The id is:  3XtlC0Z2N98MTeywSDES . Please visit the application to further document your model."
>
```

**Figure 7.** In-house R client for custom PBPK modelling-Call of the deployment function.

Once the deployment procedure is completed, a user can access the model through the Jaqpot platform and run simulations, e.g. generate forward and reverse dosimetry scenarios, provided that she/he has access to a model through the organization she/he is part of.

The model environment comprises 4 tabs: 'Overview', 'Features', 'Predict/Validate' and 'Discussion'. The 'Overview' tab provides a coarse description of the PBPK model, as well as specific directions which refer to the model, e.g. how to fill in the input section (Figure 8). The 'Features' tab informs the users about the dependent and independent features; each feature comes with description and units (Figure 9).

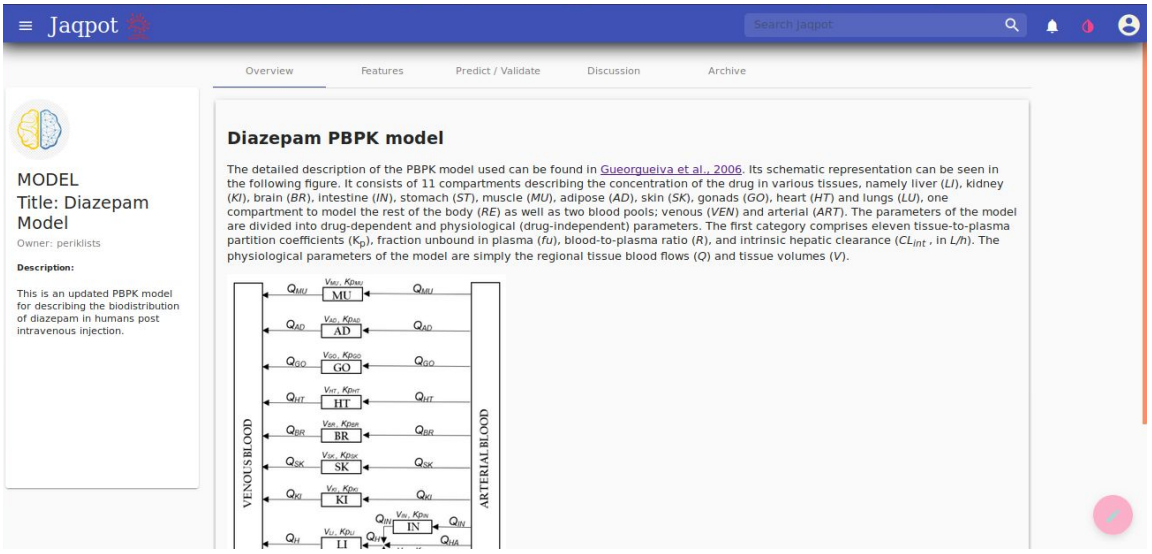



**Figure 8:** 'Overview' tab of the Diazepam model.

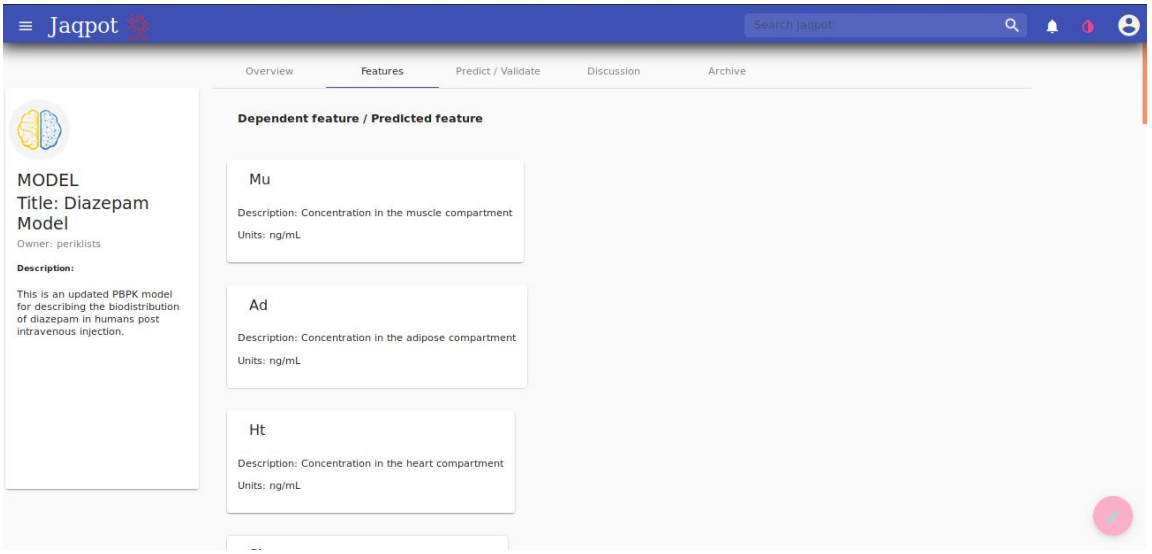



**Figure 9.** 'Features' tab of the Diazepam model.

The 'Predict/Validate' tab is the core of the model environment, where the user can provide an instance of the independent features and acquire model predictions, which, in the case of PBPK models consist mainly or concentration or mass - time profiles (Figure 10). The user can provide input in two ways: the first one is through uploading a csv file containing the respective information and the second one is through filling in the input directly in Jaqpot's Graphical User Interface (GUI). Finally, the user can add comments and remarks or ask a question regarding the model under the 'Discussion' tab (FIgure 11).

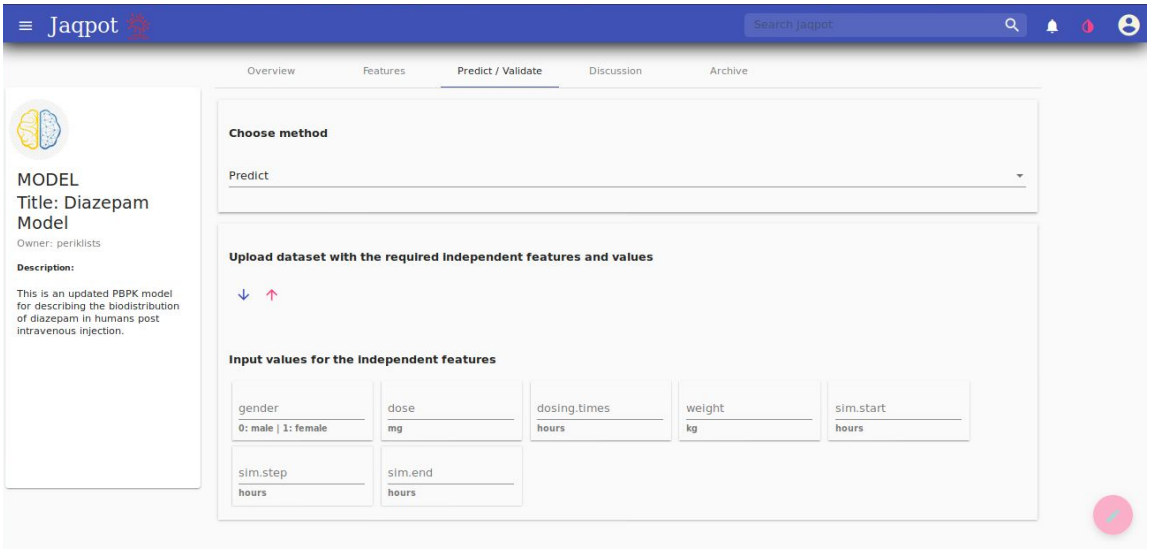



**Figure 10.** 'Predict/Validate' tab of the Diazepam model.

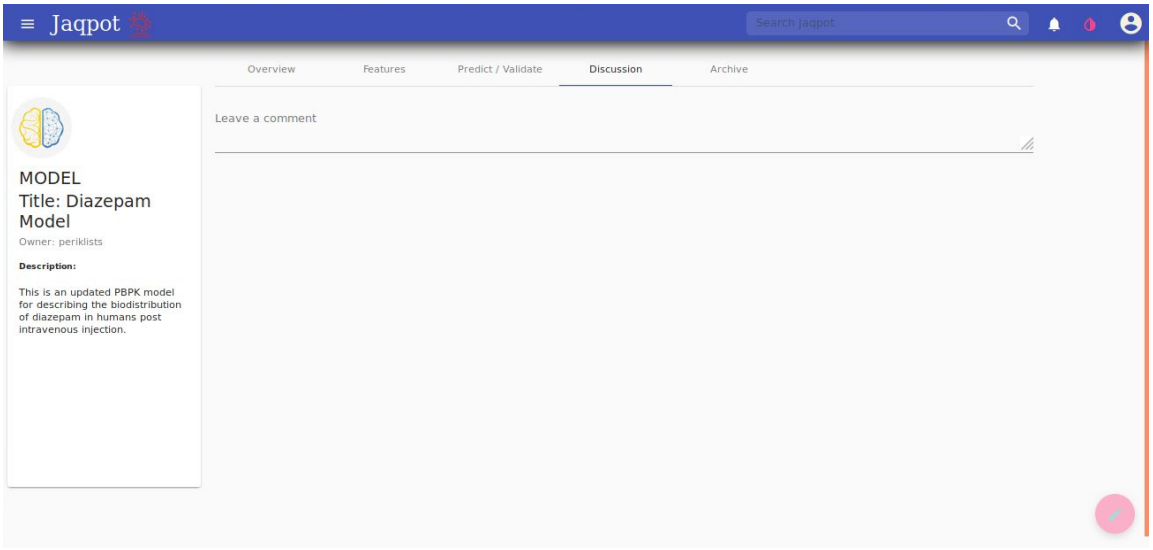



**Figure 11.** 'Discussion' tab of the Diazepam model.

More details on how to access and use PBPK models through the Jaqpot API are provided in the Appendix. Two examples of custom PBPK models that have been deployed as web services and can be accessed through the Jaqpot platform are described here. The first one is the PBPK model for diazepam in humans described in (Gueorguieva et al., 2006), which describes the biokinetics of diazepam in humans. The second example is a generic PBTK model for fish described in (Grech et al., 2019). The two models can be accessed in https://ui-jaqpot.prod.openrisknet.org/model/qof7CZIajxBHb6fU7SJz and https://ui-jaqpot.prod.openrisknet.org/model/V92BiEXxeoP35R4Nyp2y respectively.

## PBPK model for diazepam in humans

The detailed description of the PBPK model used can be found in Gueorgueiva et al. (2006). Its schematic representation can be seen in Figure 12. It refers to intravenous injection of diazepam in healthy adults and consists of 11 compartments describing the concentration of the drug in various tissues, namely liver (*LI*), kidney (*KI*), brain (*BR*), intestine (*IN*), stomach (*ST*), muscle (*MU*), adipose (*AD*), skin (*SK*), gonads (*GO*), heart (*HT*) and lungs (*LU*), one compartment to model the rest of the body (*RE*) as well as two blood pools; venous (*VEN*) and arterial (*ART*). The parameters of the model are divided into drug-dependent and physiological (drug-independent) parameters. The first category comprises eleven tissue-to-plasma partition coefficients ($K_p$), fraction unbound in plasma (*fu*), blood-to-plasma ratio (*R*), and intrinsic hepatic clearance ($CL_{int}$ , in *L/h*). The physiological parameters of the model are simply the regional tissue blood flows (*Q*) and tissue volumes (*V*).

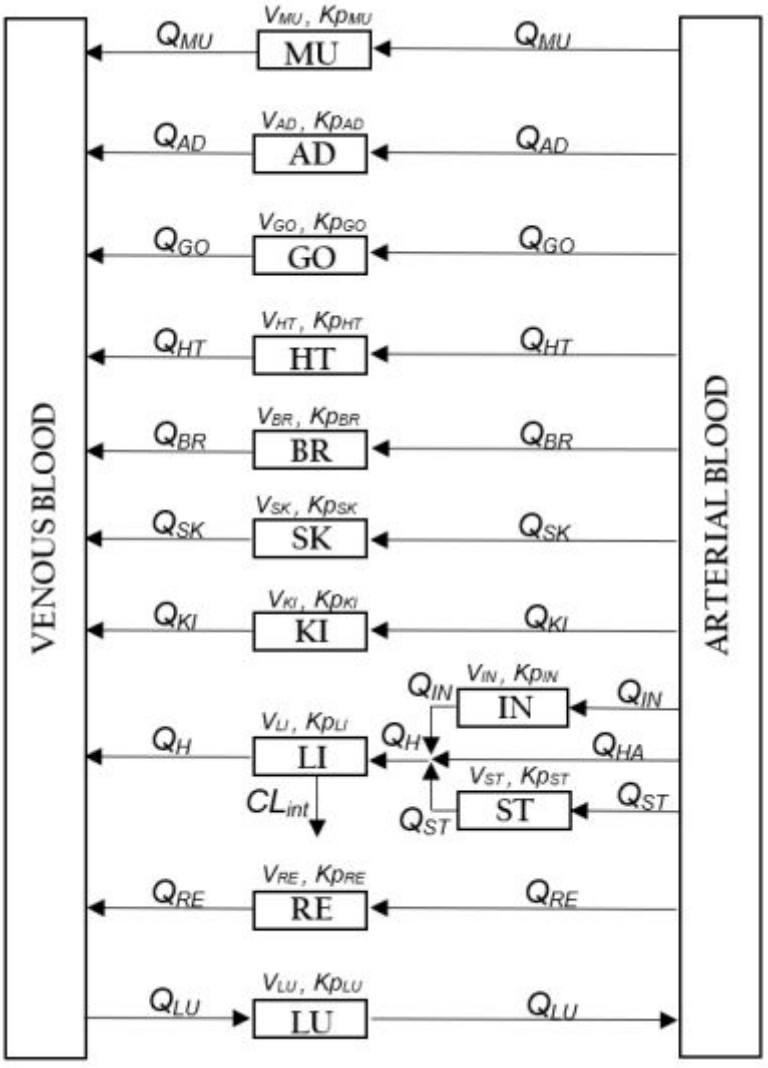



**Figure 12.** Schematic presentation of the diazepam structural model.

The values of the parameters are the product of research work performed by partners at NTUA and INERIS, involving the recalibration of the model under a different statistical model, published (this work has been published by OpenRiskNet partners (Tsiros et al., 2019)). It has been made available on Jaqpot at https://ui-jaqpot.prod.openrisknet.org/model/qof7CZIajxBHb6fU7SJz .

## Generic PBTK model for four species of fish

### Description of the model

The fish PBTK model developed by Grech et al. (2019) (Figure 13) was uploaded to Jaqpot and me available at https://ui-jaqpot.prod.openrisknet.org/model/V92BiEXxeoP35R4Nyp2y. This model describes the kinetics of xenobiotics in four species of fish, rainbow trout, fathead minnow, zebrafish and three-spined stickleback, according to a single general structure. The model comprises 12 compartments: arterial and venous blood, gills, gastrointestinal tract, skin, kidney, fat, liver, gonads, brain, poorly perfused tissues, and richly-perfused tissues. All the organs/tissues are modelled as well-mixed compartments with a blood flow-limited distribution.

Both gastro-intestinal (in case of food ingestion) and branchial absorption are modelled. Chemical binding to plasma proteins is considered by introducing an unbound fraction of the chemical in plasma. Metabolism is modelled in the liver or in the plasma. Excretion can occur via urine, expired water, and faeces, as unabsorbed fraction or by biliary excretion. Compounds excreted by the gills and urine are released in the water and can be reabsorbed in static water conditions.

The PBTK model includes a growth model based on the dynamic energy budget (DEB) theory (Kooijman, 2010). Two growth models were implemented. The first is a standard DEB model (Kooijman, 2010) and can be represented by a von Bertalanffy growth curve. This model was used for zebrafish and stickleback, in accordance with the Add-my-pet database (www.bio.vu.nl/thb/deb/deblab/add_my_pet). The second growth model is described by a DEB model with type M acceleration (Kooijman and Lika, 2014) and is characterised by an up-curving of length-at-time since birth at constant food (Kooijman, 2014). This model was used for rainbow trout and fathead minnow.

In case the tissue:blood partition coefficients are unknown, they can be automatically estimated by QSAR modelling. However, the user has to ensure that the chemical studied in the applicability domain of the QSAR models that are implemented (non polar, non ionizable, and log Kow between 2 and 6).

An extensive literature search was performed to identify experimental data informing the physiological PBTK model's parameters for the four species, distinguishing male and female when data was available.

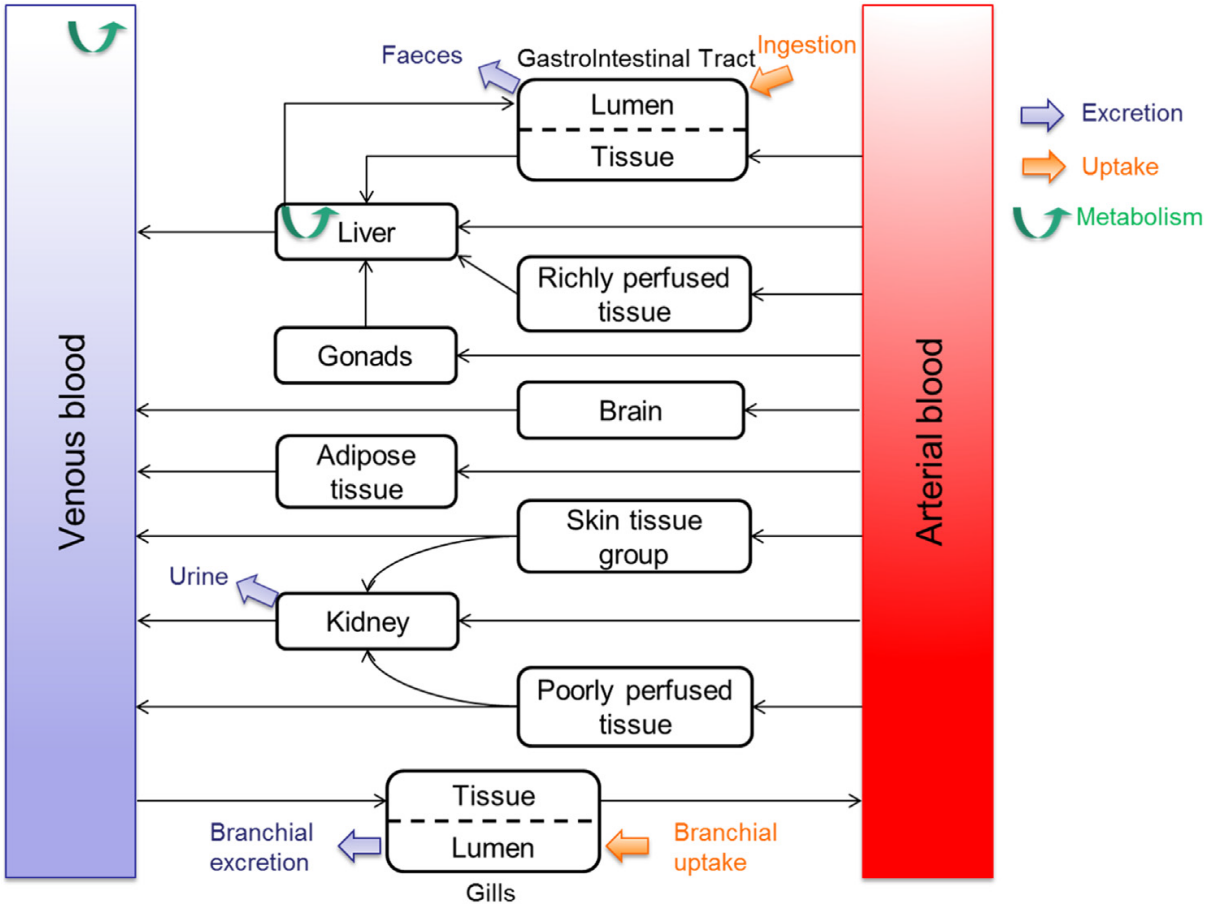



**Figure 13.** PBTK model structure for fish.

### Implementation in Jaqpot

Because this PBTK model is generic, the user defines as many as 40 parameters which relate to:

- Chemical-specific absorption: Ku, frac_abs (oral), PC_bw (if NA estimated using log_Kow),
- Chemical-specific excretion: Ke_bile, K_BG, Ke_urine, Ke_feces, PC_bw, unbound_fraction
- Chemical-specific distribution: PC_l, PC_k, PC_s, PC_f, PC_p, PC_r, PC_gi, PC_go, PC_b, Ratio_blood_plasma
- Chemical-specific metabolism: rate_plasma, Cl_liver, Km, Vmax
- Exposure and experimental setup: WaterQuantity, IngestQuantity, IVQuantity, V_water, Ke_water, period, frac_renewed, time_first_exposure, time_last_exposure, Bw_i, Temperature, f_cst, Gender, Species

- Simulations to be carried out: sim.start, sim.end, sim.step

The output of the simulations is a table of internal concentrations (C_art, C_plasma, C_liver, C_kidney, C_brain, C_fat, C_gonads, C_GIT, C_rp, C_pp, C_skin, C_tot), Body weight (BW), fish length, and total amount administered to/absorbed by the fish, as a function of time.

# Integration of PKsim in OpenRiskNet Infrastructure through the Jaqpot modelling platform

This subsection presents a step-by-step demonstration on how to create a PBPK model using PKSim (Willmann et al., 2003) and expose it as a web service through the Jaqpot modelling platform.

Software requirements: The user should download and install in his/her PC the Open Systems Pharmacology Suite:
https://github.com/Open-Systems-Pharmacology/Suite/releases/tag/v7.1.0

and the gene expression database:

https://github.com/Open-Systems-Pharmacology/Suite/releases/download/v7.1.0/GENEDB_human.mdb

In PKSim select Utilities, then Options, then “Application” and for Human you give the path where the database is stored.

All other necessary files that should be downloaded are available in the google drive folder used for the OpenRiskNet/OpenTox Euro 2017 Biokinetics Workshop, where this example was presented:

https://drive.google.com/drive/folders/1wGmqNYI8GnDL_orrE2JqPQAMauHStbPj

Assuming that the PBPK model has been saved in xml format (as explained in Powerpoint presentation), the following steps should be followed in Swagger JaqPot API documentation http://jaqpot.org:8080/jaqpot/swagger/ :

1) Create a JaqPot dataset containing the physiological parameters of the individual on which the PBPK model was developed. We assume that these parameters (age, height, weight) are included in a csv file testPK.csv
    a. Use the **POST /dataset/createDummyDataset** method: Choose the ***testPK.csv*** file and give any title and description to the produced dataset. In the end of the response body a dataset id is generated.
    b. Use the **GET /dataset/{id}** method: Copy the dataset id in the relevant field and press Try it out!. In the Request URL the full dataset URI of the produced dataset is shown. Store the URI of the dataset.
    Example: http://jaqpot.org:8080/jaqpot/services/dataset/0JRSR55QrpHpMi (the dataset can be accessed through Swagger using its dataset ID, **0JRSR55QrpHpMi**).

2) Produce a JaqPot resource that implements the PBPK model as a web service
    a. Use the **POST /biokinetics/train** method. Insert the following information in the parameter fields:
        - *File*: Choose the PBPK xml file ***sim_individual_1_XML.xml***
        - Dataset-uri: Copy the full dataset URI
        - Title and description: Any title and description is fine
        - Algorithm-uri: http://jaqpot.org:8080/jaqpot/services/algorithm/pk-sim (more information on this algorithm can be obtained over Swagger using the **GET /algorithm/{id}** method http://jaqpot.org:8080/jaqpot/swagger/#!/algorithm/getAlgorithm with the ID of the algorithm, **pk-sim**)
        - *Parameters*: {"ageUnit":["years"], "individual":[1],"heightUnit":["m"], "weightUnit":["kg"],"drug":["Theophylline"]}

      After pressing the Try it out! button, a task ID is generated In the end of the response body
    b. Use the **GET /task/{id}** method. Copy the task ID in the relevant field and press the Try it out! Button. If the status in the end of the response body is not COMPLETED wait for a few seconds and try again. The model has been created and its complete URI is shown in the response body. You can now share the model with the rest of the world. For the next step store the model id. This is the id after model/ in “result” in the response body
3) Use the model for obtaining drug concentration-time profiles
    a. Use the **POST /model/{id}** method. In the dataset_uri you can use the same dataset URI that was used in the first step. Alternatively, you can create an alternative dataset for a different individual. In the id field copy the id of the produced model. After pressing the Try it out! button a task ID is created.
    b. Use the **GET /task/{id}** method. Copy the task ID in the relevant field and press the Try it out! Button. The dataset with drug concentration –time data has been created. For the next step store the dataset id. This is the id after dataset/ in “result” in the response body
    c. Use the **GET /dataset/{id}** method: Copy the dataset id in the relevant field and press Try it out!. In the Response Body, drug concentration values are shown along with corresponding time points.

# Reverse dosimetry

## Workflow for reverse dosimetry

Reverse dosimetry of diazepam in humans was performed by selecting a relevant internal concentration from the literature. The diazepam PBPK model (at https://ui-jaqpot.prod.openrisknet.org/model/qof7CZIajxBHb6fU7SJz) was then run using forward dosimetry (i.e. predicting the internal concentration consequently to a given exposure scenario) iteratively in order to estimate the exposure dose which produces the selected internal concentration (Figure 14).

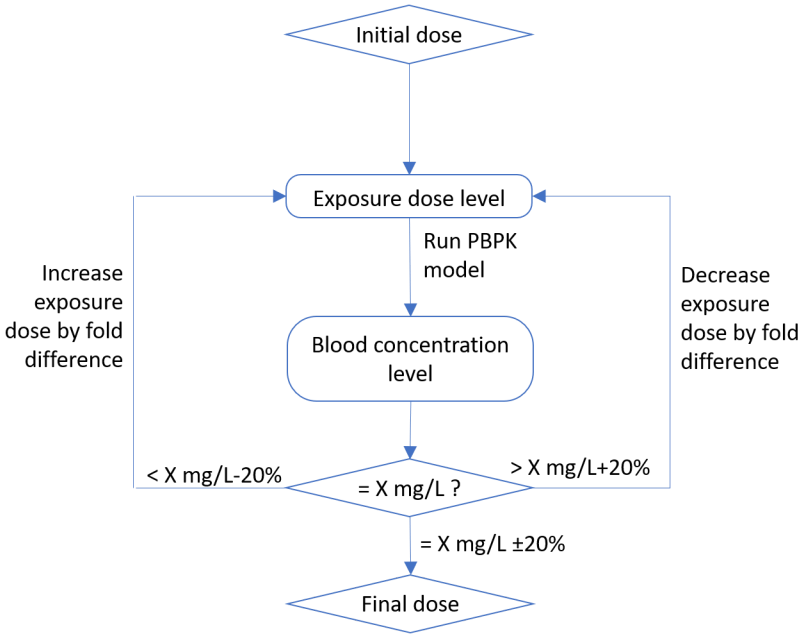



**Figure 14.** Workflow for reverse dosimetry of diazepam. X is the target internal concentration.

### Diazepam

No biomonitoring of diazepam in the general population was identified in the literature. Summary data on diazepam concentrations in drivers apprehended for driving under the influence of drugs in Sweden was available in the literature (Jones et al., 2012). A median value of 0.20 mg/L diazepam in blood was estimated in 1,000 blood samples where diazepam and its metabolite nordiazepam were both present.

The diazepam model simulates intravenous injections. The exposure scenario was assumed to be daily injections. The initial daily dose in the reverse dosimetry is set to 5mg (equivalent to one pill per day).

### Chlorpyrifos

Reverse dosimetry with the fish PBTK model (at https://ui-jaqpot.prod.openrisknet.org/model/V92BiEXxeoP35R4Nyp2y) was performed

with the same procedure in the case of continuous waterborne exposure to chlorpyrifos in rainbow trout (Grech et al., 2019). PBPK parameter values are listed in Table 1.

**Table 1.** User-defined parameter values used for the chlorpyrifos PBTK in rainbow trout.

| Parameter | Unit | Value | Source |
|---|---|---|---|
| Log $K_{ow}$ | - | 4.96 | |
| $V_{max}$ | µg/d/g liver | 3375.38 | Lavado and Schlenk, 2011 |
| $K_m$ | µg/mL | 50.31 | Lavado and Schlenk, 2011 |
| Cl_liver | mL.d-1 | NA | |
| Rate_plasma | | 0 | |
| Unbound_fraction | - | 0.05 | Weelling et al., 1992 |
| Ke urine | $d^{-1}$ | 0 | Weelling et al., 1992 |
| Ke bile | $d^{-1}$ | 0 | Weelling et al., 1992 |
| Ke feces | $d^{-1}$ | 0.83 | Nichols et al., 2004 |
| $K_{BG}$ | $d^{-1}$ | 1e12 | |
| Ku | $d^{-1}$ | 0 | |
| Frac_abs | | 0 | |
| Ratio blood to plasma | - | 1 | |
| Temperature* | °C | 17 | |
| Ke water* | $d^{-1}$ | 0 | |
| WaterQuantity* | µg | 9e09 | |
| V_water* | mL | 1e12 | |
| IngestQuantity* | µg | 0 | |
| IVQuantity* | µg | 0 | |
| Period* | d | NA | |
| Time_first_dose* | d | NA | |
| Time_last_dose* | d | NA | |
| Frac_renewed* | | 0 | |
| BW_i* | g | 100 | |
| f_cst* | | 0.3 | |
| Species* | | RT | |
| Gender* | | 0 | |

*Study dependent

**Table 2.** QSAR-estimated partition coefficients

| Partition coefficient | Value |
|---|---|
| blood:water | 1119 |
| liver:blood | 8.56 |
| gonads:blood | 17.93 |
| brain:blood | 5.95 |
| fat:blood | 8.15 |
| skin:blood | 3.75 |
| GIT:blood | 6.03 |
| kidney:blood | 13.54 |
| RPT:blood | 5.52 |
| PPT:blood | 2.04 |

No biomonitoring data on internal chlorpyrifos concentrations in fish was available. Reverse dosimetry was performed in order to estimate the chlorpyrifos concentration level in water that would elicit the EC50 level for the EROD biomarker on rainbow trout liver cells (RTL-W1) in vitro which was estimated to be 0.022 mg/L (Babín et al., 2005) The initial water concentration in the reverse dosimetry procedure was set to 9 µg/L, which is the lower range of LC50 in rainbow trout (Wheelock et al., 2005) determined by Phipps and Holcombe (Phipps and Holcombe, 1985). The higher range of LC50 values is 45 µg/L for 96hr acute toxicity (Kikuchi et al., 1996).

## Implementation in Jaqpot

### Diazepam

For the diazepam PBPK, the parameters for repeated exposure can be entered in the fields (Figure 15).

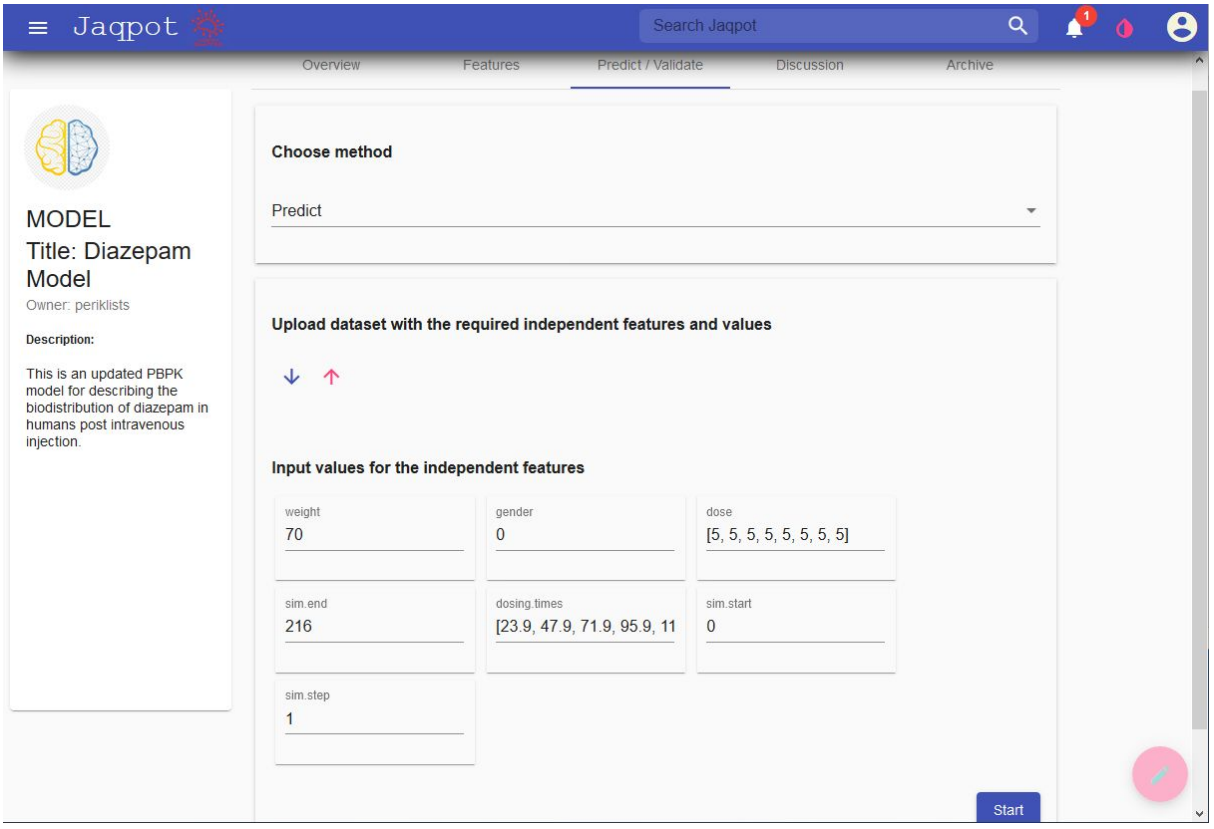



**Figure 15.** Parameters for the diazepam PBPK with repeated exposure.

### Chlorpyrifos

For the fish PBTK model, the parameters for waterborne exposure to chlorpyrifos can also be entered either as a csv file or in the parameter input fields (Figure 16 and Figure 17).

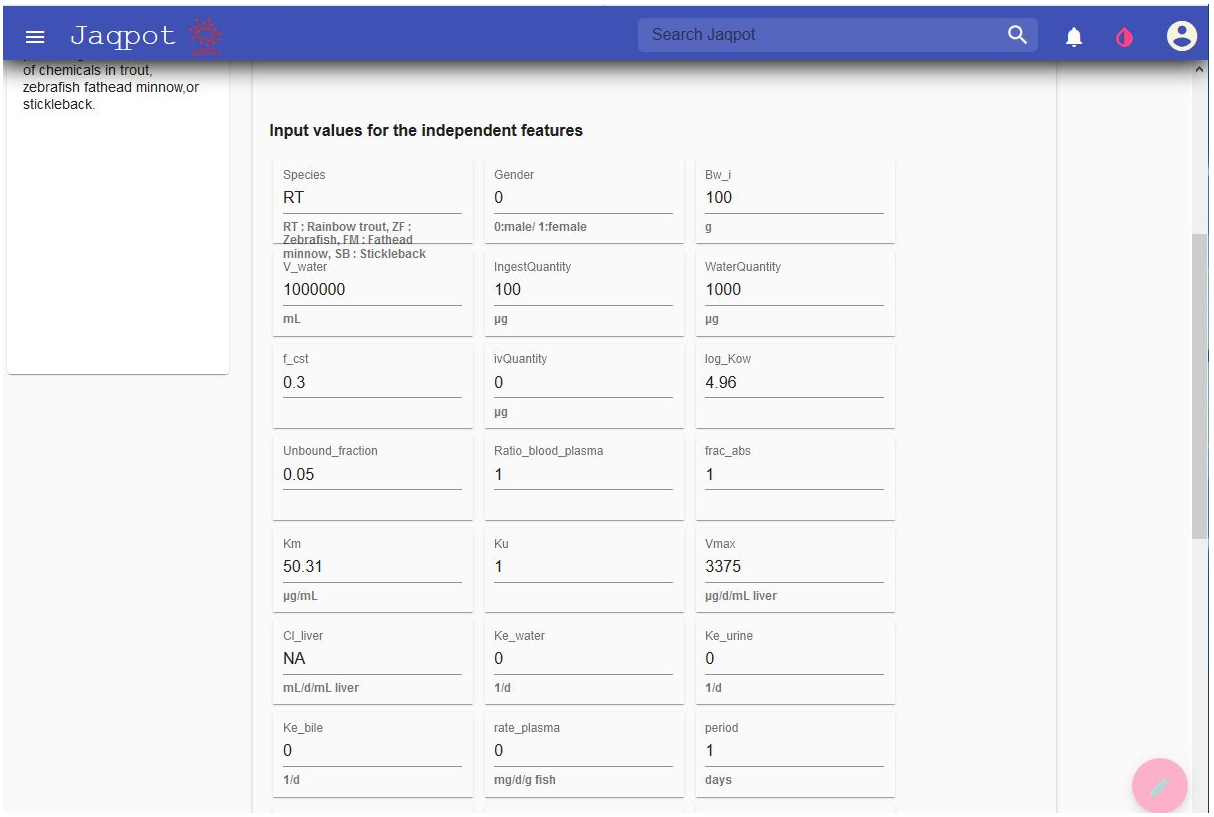



**Figure 16.** Parameters for the fish PBTK for chlorpyrifos.

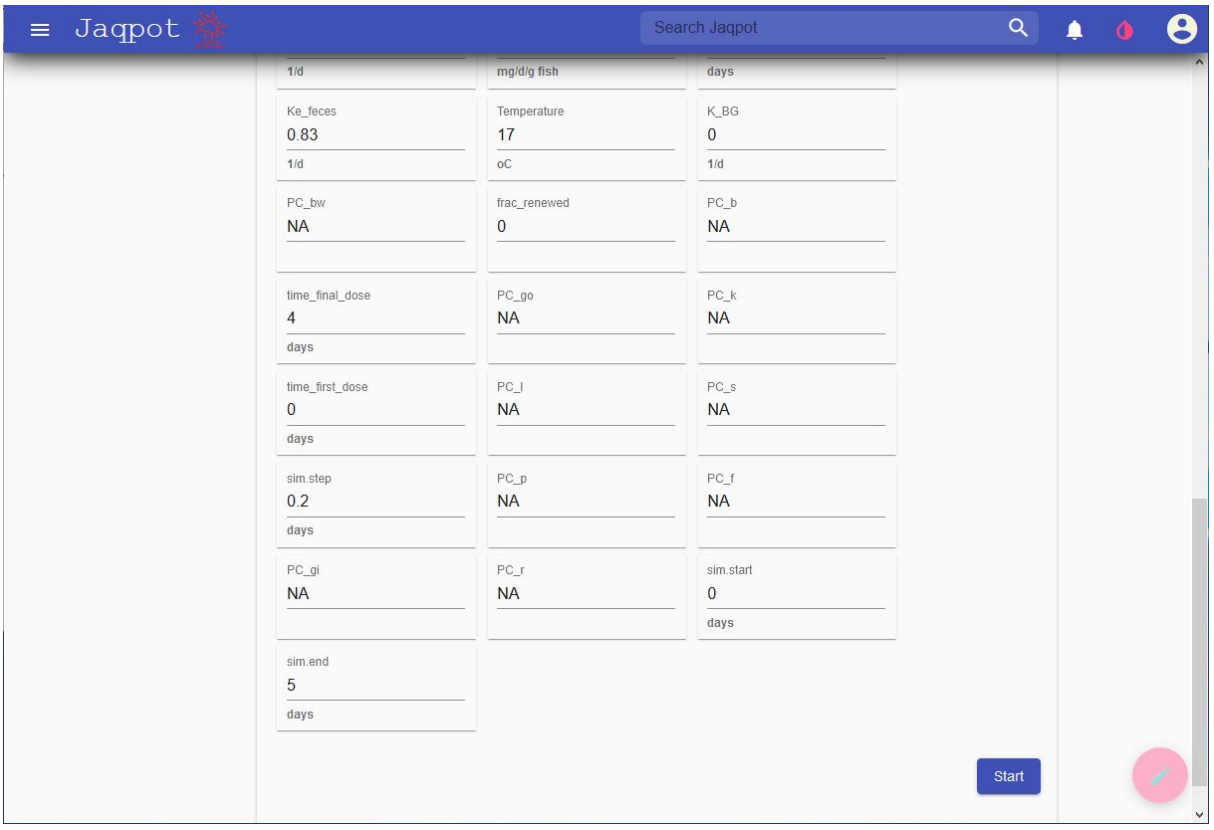



**Figure 17.** Parameters for the fish PBTK (continued).

## Results

### Diazepam

Diazepam levels in blood were predicted in a 70kg male subject first with a 5mg IV daily dose (Figure 18 and Figure 19). the level increased sharply, due to the simulated injection, then decreased sharply and followed a slow, approximately linear, decrease one hour after the dose and until the next dose. Steady state was obtained after 7 doses, when the peak and lowest concentration levels were the same every day.

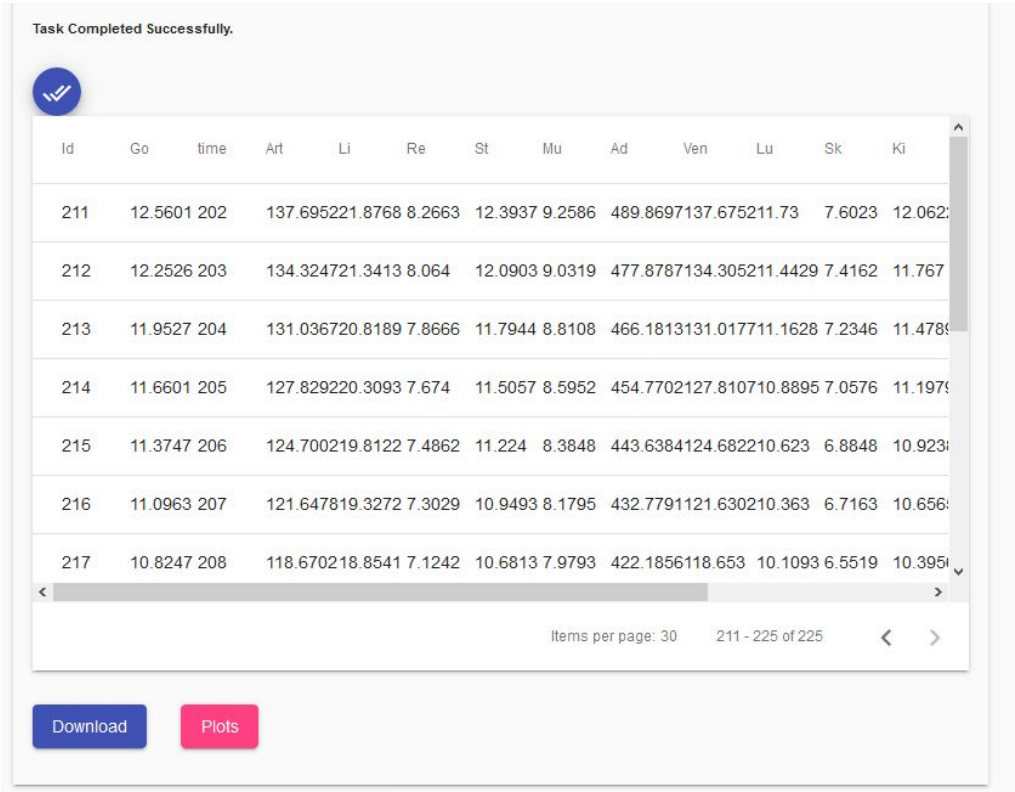



**Figure 18.** Output at last simulation times for repeated exposure to diazepam - mg daily IV doses in 70 kg male subjects.

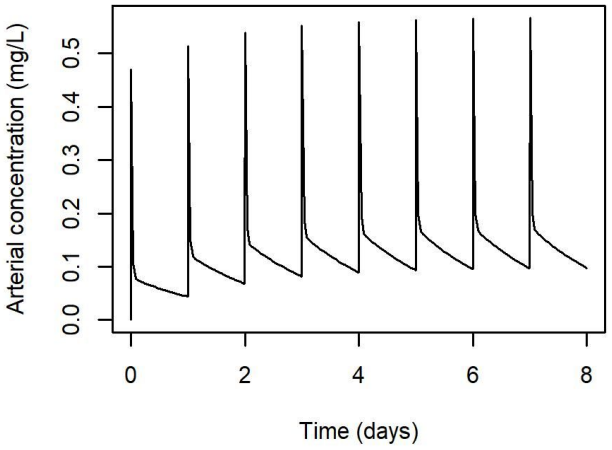



**Figure 19.** Predicted diazepam concentrations in blood resulting from 5mg daily IV doses in 70 kg male subjects.

The strong peak in concentration could be due to the simulated mode of administration (IV). Since diazepam would most often be administered orally, the time point selected for

comparison with the biomonitoring data in drivers was 4 hours after the dose, well after the peak concentration.

4 hours after the 8th dose, the concentration in blood was 0.159 mg/L, which is very close to the median in drivers (0.20 mg/L). The simulated administered dose was increased to 6.26 mg according to the workflow. The resulting blood concentration 4 hours after the 8th dose was 0.20 mg/L (Figure 20).

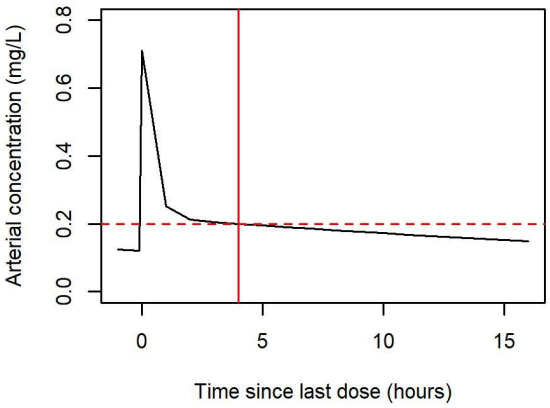



**Figure 20.** Predicted diazepam concentrations in blood resulting from 6.26 mg daily IV doses in 70 kg male subjects (focus on period after 8th dose).

### Chlorpyrifos

Exposure of rainbow trout to a constant exposure chlorpyrifos was simulated for juvenile fish (body weight set to 100 g) at 17°C in order to reproduce the experimental conditions reported by (Phipps and Holcombe, 1985). Internal concentrations reach steady state after about 60 days when fish are exposed to 9 µg/L (Figure 21A). After 96 hours of simulated exposure to chlorpyrifos, as expected the lethality LC50 resulted in liver concentrations which where higher than the EC50 for the EROD biomarker. Indeed, the liver concentration reached 21 mg/L, which is 1000-fold higher than the 0.022 mg/L in vitro EC50 for EROD inhibition. Consequently, to perform reverse dosimetry, the exposure concentration was set to 0.0093 µg/L which resulted in a liver concentration of 0.021 after 96 hours (Figure 21B), which was within 20% of the target internal concentration.

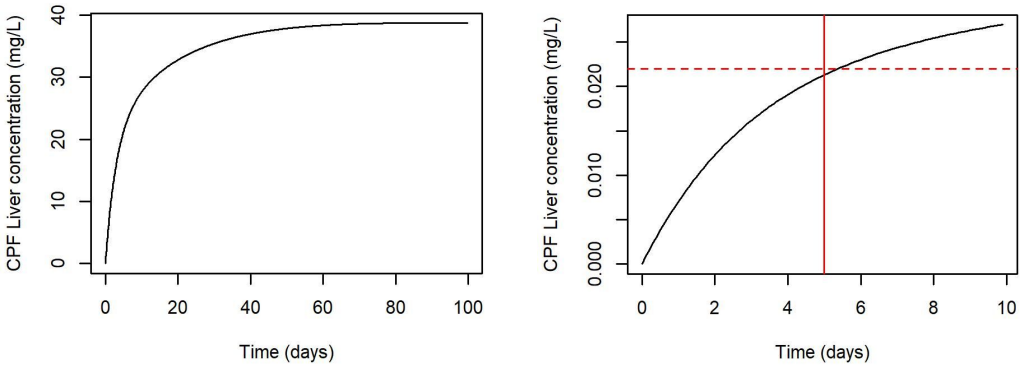



**Figure 21.** Predicted concentrations of chlorpyrifos in rainbow trout liver exposed to 9 µg/L (A) or 0.0093 µg/L (B) chlorpyrifos at 17°C.

In both diazepam and chlorpyrifos examples, reverse dosimetry was straightforward due to almost linear kinetics in the dose range considered.

# REFERENCES


Babín, M.M. and Tarazona, J.V., *In vitro toxicity of selected pesticides on RTG-2 and RTL-W1 fish cell lines.* Environmental Pollution, 2005. **135**(2): p. 267-274. https://doi.org/10.1016/j.envpol.2004.11.001

Chomenidis, C., Drakakis, G., Tsiliki, G., Anagnostopoulou, E., Valsamis, A., Doganis, P., Sopasakis, P., Sarimveis, H., *Jaqpot Quattro: A Novel Computational Web Platform for Modeling and Analysis in Nanoinformatics.* Journal of Chemical Information and Modeling, 2017. **57**: p. 2161-2172. https://doi.org/10.1021/acs.jcim.7b00223

Grech, A., Tebby, C, Brochot, C., Bois, F.Y., Bado-Nilles, A., Dorne, J.L., Quignot, N., Beaudouin, R., *Generic physiologically-based toxicokinetic modelling for fish: Integration of environmental factors and species variability.* Science of the Total Environment, 2019. **651**: p. 516-531. https://doi.org/10.1007/s10928-019-09630-x

Gueorguieva, I., Aarons, L., Rowland, M., *Diazepam Pharamacokinetics from Preclinical to Phase I Using a Bayesian Population Physiologically Based Pharmacokinetic Model with Informative Prior Distributions in Winbugs*. Journal of Pharmacokinetics and Pharmacodynamics, 2006. **33**(5): p. 571–594.

Jones, A. and Holmgren, A., *Concentrations of Diazepam and Nordiazepam in 1000 Blood Samples From Apprehended Drivers —Therapeutic Use or Abuse of Anxiolytics?* Journal of Pharmacy Practice, 2012. **26**(3): p. 198-203. https://doi.org/10.1177/0897190012451910

Jones, H.M. and Rowland-Yeo, K., *Basic concepts in physiologically based pharmacokinetic modeling in drug discovery and development*. CPT Pharmacometrics Syst Pharmacol, 2013. **2**:e63. doi: https://doi.org/10.1038/psp.2013.41

Kikuchi, M., Miyagaki, T. and Wakabayashi, M., *Evaluation of Pesticides Used in Golf Links by Acute Toxicity Test on Rainbow Trout.* NIPPON SUISAN GAKKAISHI, 1996. **62**(3): p. 414-419. https://doi.org/10.2331/suisan.62.414

Kilford, P. J., Gertz, M., Houston, J. B. and Galetin, A., *Hepatocellular binding of drugs: correction for unbound fraction in hepatocyte incubations using microsomal binding or drug lipophilicity data*. Drug Metabolism and Disposition, 2008. **36**(7): p. 1194-7. https://doi.org/10.1124/dmd.108.020834

Kooijman, S.A., *Notation of dynamic energy budget theory for metabolic organisation*. 2010: Cambridge University Press.

Kooijman, S.A.L.M., *Metabolic acceleration in animal ontogeny: An evolutionary perspective.* Journal of Sea Research, 2014. **94**: p. 128-137. https://doi.org/10.1016/j.seares.2014.06.005

Kooijman, S.A.L.M. and Lika, K., *Resource allocation to reproduction in animals*. Biological Reviews, 2014. **89**(4): p. 849-859. https://doi.org/10.1111/brv.12082

Lavado, R. and Schlenk, D., *Microsomal biotransformation of chlorpyrifos, parathion and fenthion in rainbow trout (Oncorhynchus mykiss) and coho salmon (Oncorhynchus kisutch): Mechanistic insights into interspecific differences in toxicity.* Aquatic Toxicology, 2011. **101**(1): p. 57-63. https://doi.org/10.1016/j.aquatox.2010.09.002

Nestorov, I., *Whole body pharmacokinetic models*. Clin Pharmacokinet, 2003. **42**(10): p. 883-908. https://doi.org/10.2165/00003088-200342100-00002

Nestorov, I.A., Aarons, L.J., Arundel, P.A., Rowland, M., *Lumping of whole-body physiologically based pharmacokinetic models*. J Pharmacokinet Biopharm, 1998. **26**(1): p. 21–46. doi: https://doi.org/10.1023/A:1023272707390

Nichols, J.W., Fitzsimmons, P.N., Whiteman, F.W., Dawson, T.D., Babeu, L., Juenemann, J., *A physiologically based toxicokinetic model for dietary uptake of hydrophobic organic*

*compounds by fish - I. Feeding studies with 2,2 ',5,5 '-tetrachlorobiphenyl.* Toxicological Sciences, 2004. **77**(2): p. 206-218. https://doi.org/10.1093/toxsci/kfh033

Ooms, J. *The OpenCPU System : Towards a Universal Interface for Scientific Computing through Separation of Concerns*. *arXiv,* 2014. p. 1–23.

Pearce, R., Setzer, R., Strope, C., Sipes, N., Wambaugh, J., *httk: R Package for High-Throughput Toxicokinetics*. Journal of Statistical Software, *2017.* **79**(4): p. 1 - 26. doi:http://dx.doi.org/10.18637/jss.v079.i04

Phipps, G.L. and Holcombe, G.W., *A method for aquatic multiple species toxicant testing: Acute toxicity of 10 chemicals to 5 vertebrates and 2 invertebrates.* Environmental Pollution Series A, Ecological and Biological, 1985. **38**(2): p. 141-157. https://doi.org/10.1016/0143-1471(85)90073-X

Pilari, S., Huisinga, W., *Lumping of physiologically-based pharmacokinetic models and a mechanistic derivation of classical compartmental models*. J Pharmacokinet Pharmacodyn, 2010. **37**(4): p. 365–405. doi: https://doi.org/10.1007/s10928-010-9165-1

Sturm, R.A., *Computer Model for the Clearance of Insoluble Particles from the Tracheobronchial Tree of the Human Lung*. Comput Biol Med, 2007. **37**(5): p. 680–690. doi: https://doi.org/10.1016/j.compbiomed.2006.06.004

Tsiros, P., Bois, F. Y., Dokoumetzidis, A., Tsiliki, G., Sarimveis, H., *Population pharmacokinetic reanalysis of a Diazepam PBPK model: a comparison of Stan and GNU MCSim.* Journal of Pharmacokinetics and Pharmacodynamics, 2019. 46(2): p. 173–192. https://doi.org/10.1016/0143-1471(85)90073-X

Weelling, W. and De Vries, J., *Bioconcentration kinetics of the organophosphorus insecticide chlorpyrifos in guppies (Poecilla reticulata).* Ecotoxicology and Environmental Safety, 1992. **23**: p. 64–75. https://doi.org/10.1016/0147-6513(92)90022-U

Wheelock, C.E., et al., *Individual variability in esterase activity and CYP1A levels in Chinook salmon (Oncorhynchus tshawyacha) exposed to esfenvalerate and chlorpyrifos.* Aquatic Toxicology, 2005. **74**(2): p. 172-192. https://doi.org/10.1016/j.aquatox.2005.05.009

Willmann S, Lippert J, Sevestre M, Solodenko J, Fois F, Schmitt W. *PK-Sim®: a physiologically based pharmacokinetic 'whole-body' model*. BIOSILICO, 2003. **1**(4): p. 121–4. http://dx.doi.org/10.1016/S1478-5382(03)02342-4.

# APPENDIX

## USER GUIDANCE ON PBPK MODELS IN JAQPOT

### HOW TO ACCESS A CUSTOM PBPK MODEL

An important module introduced in Jaqpot 5 is custom PBPK models that have been deployed through a developed R client. A Jaqpot user can run simulations using these models, e.g. generate forward and reverse dosimetry scenarios, provided that he has access to a model through the organization she/he is part of.

After logging into Jaqpot, the user is directed to Jaqpot's Home page, from where he can go to the models section by clicking the 'Models' tab on the left banner of the screen (Figure A1).

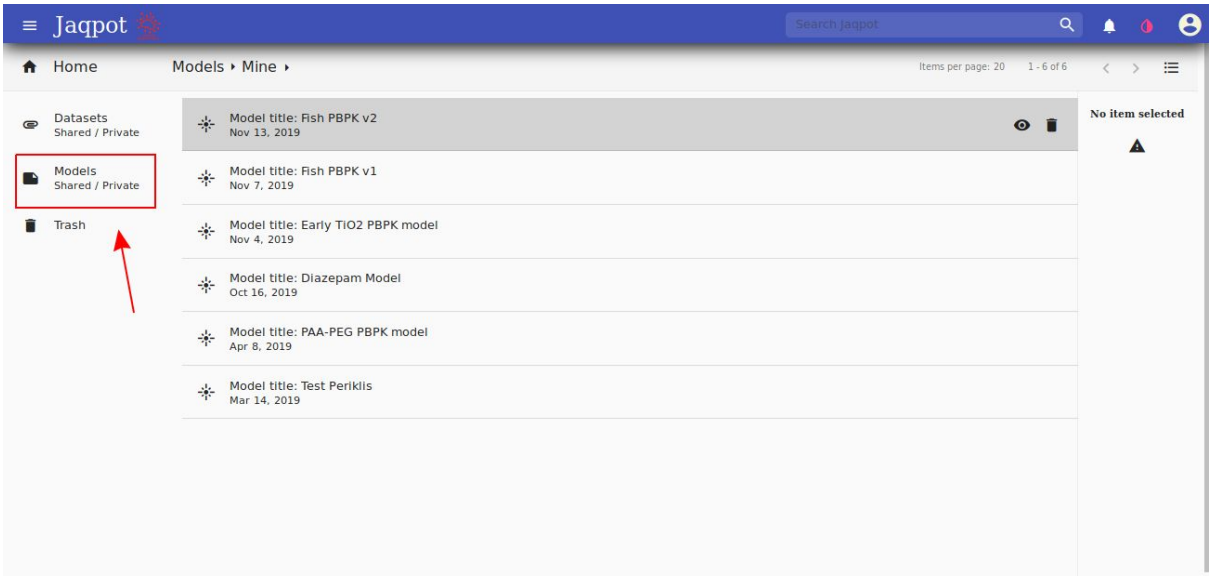



**Figure A1.** 'Models' Tab.

The initial screen includes models deployed by the user, private and shared ones. In order to access models that have been shared to the user via organisations, the user should click on the arrow right next to the 'Mine' tab on the top of the screen and then click the 'Shared' button (Figure A2).

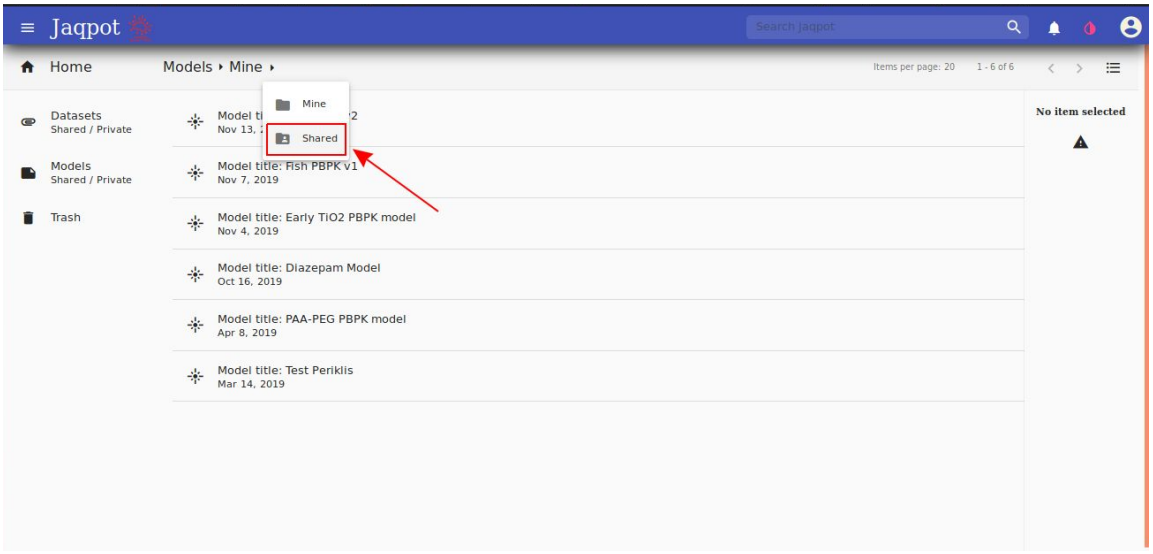



**Figure A2.** Selecting Share models.

Following that, the user can access a specific model by clicking on the 'View' button, which is located on the far right end of the model's row (Figure A3).

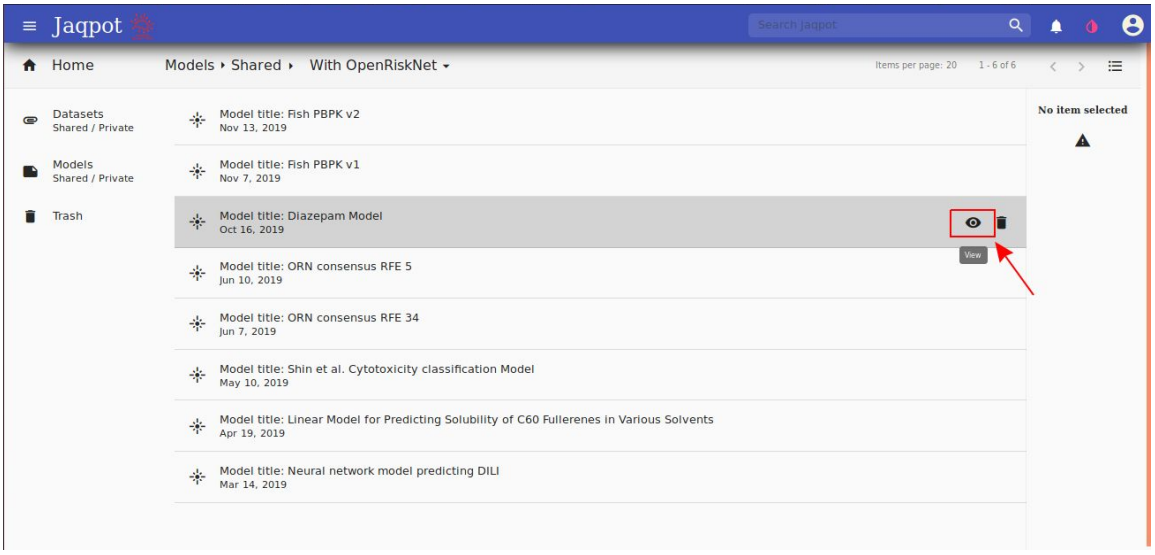



**Figure A3**. Clicking on the 'View' tab.

## NAVIGATING A MODEL PAGE

The model environment comprises 4 tabs: 'Overview', 'Features', 'Predict/Validate' and 'Discussion'. The 'Overview' tab provides a coarse description of the PBPK model, as well as specific directions which refer to the model, e.g. how to fill in the input section (Figure A4).

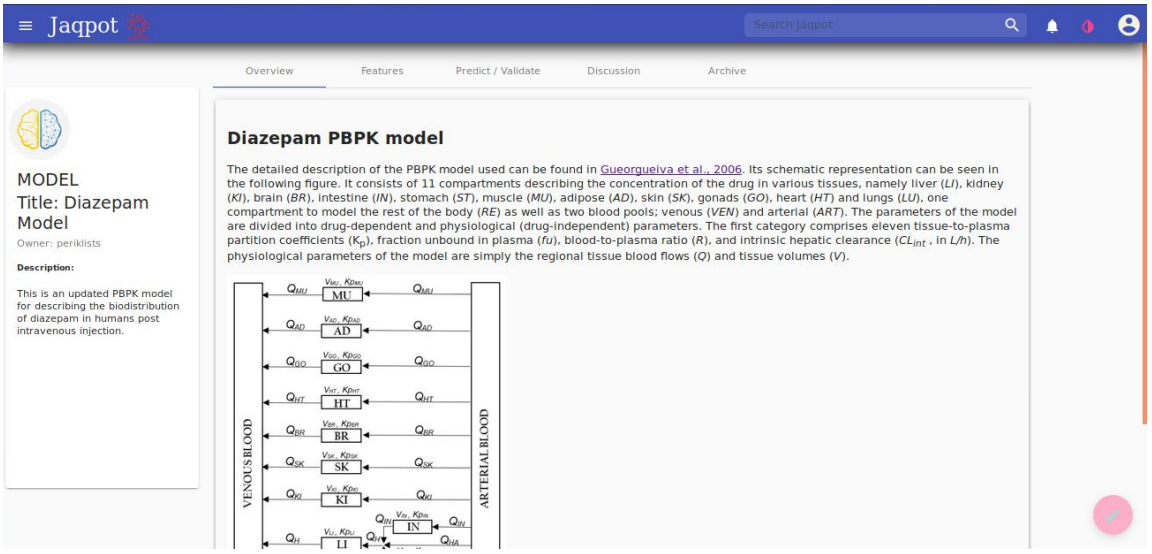



**Figure A4**. 'Overview' tab of the Diazepam model.

The 'Features' tab informs the users about the dependent and independent features; each feature comes with description and units (Figure A5).

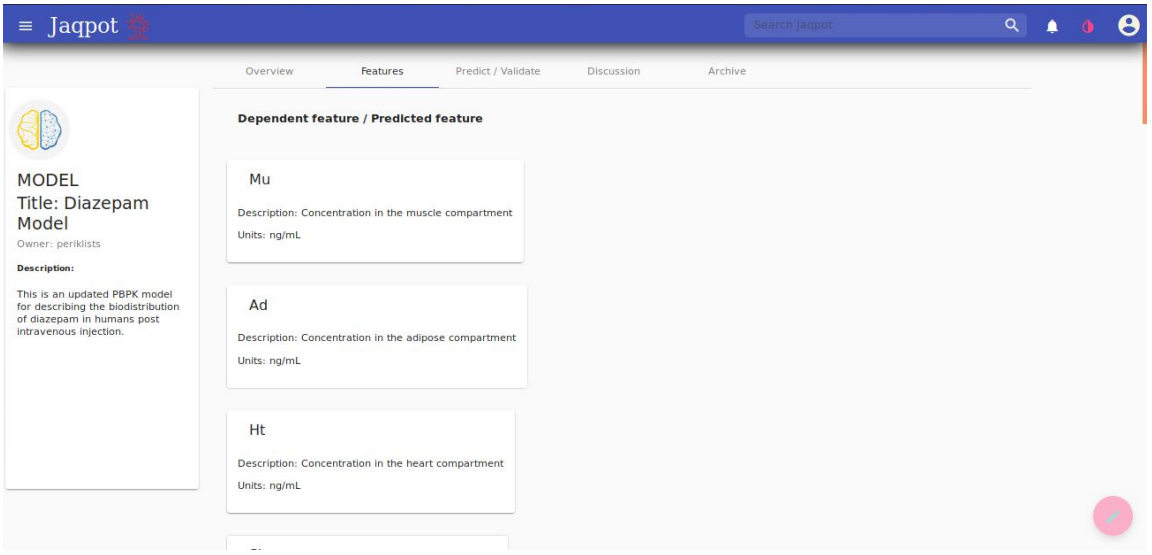



**Figure A5**. 'Features' tab of the Diazepam model.

The 'Predict/Validate' tab is the core of the model environment. Here, the user can provide an instance of the independent features and acquire the model predictions, which, in the case of PBPK models consist mainly or concentration or mass- time profiles. The user can provide the input in two ways: the first one is through uploading a csv file containing the respective information and the second one is through filling in the input directly in

Jaqpot's Graphical User Interface (GUI). In case the input consists of many features, it is strongly recommended that the user follows the first method, i.e. download the csv template (Figure A6), fill in the values (Figure A7), upload the complete csv (Figure A8), select 'None' in the pop up window asking for a dataset ID (FIgure A9) and then start the prediction process (Figure A10).

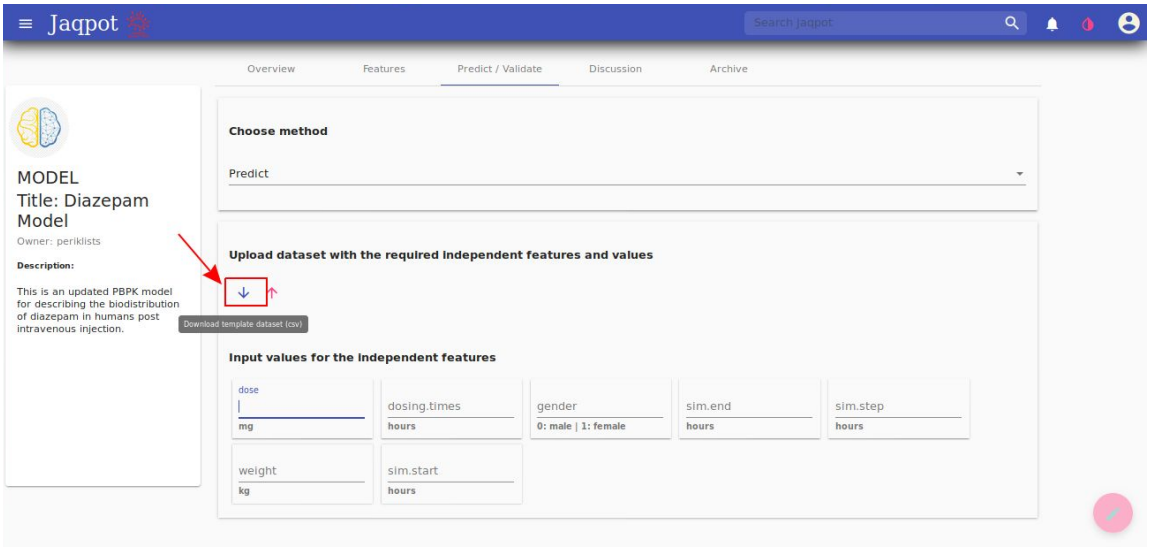



**Figure A6**. How to download a dataset template in csv format.

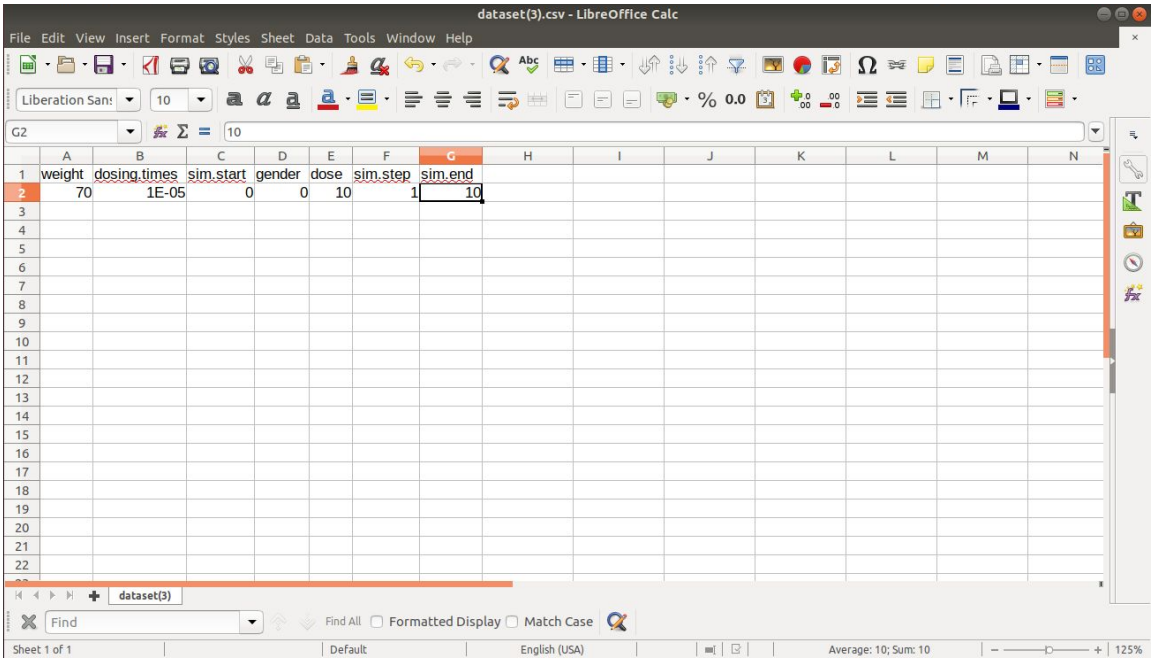



**Figure A7**. Complete the csv with appropriate values.

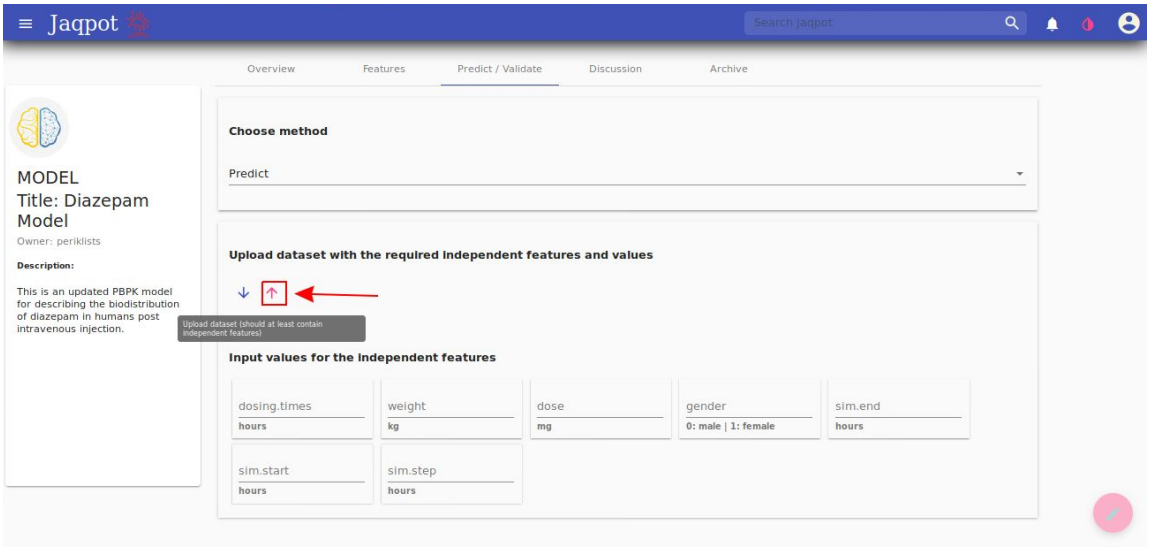



**Figure A8**. Upload the complete dataset.

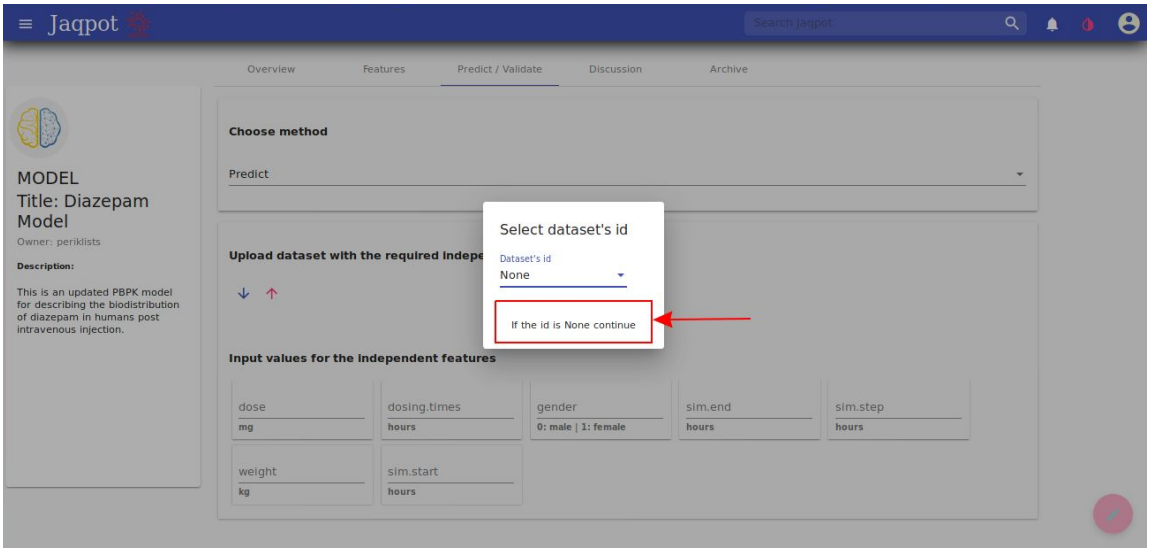



**Figure A9**. Select ‘None’ in the dataset id and then click on ‘continue’.

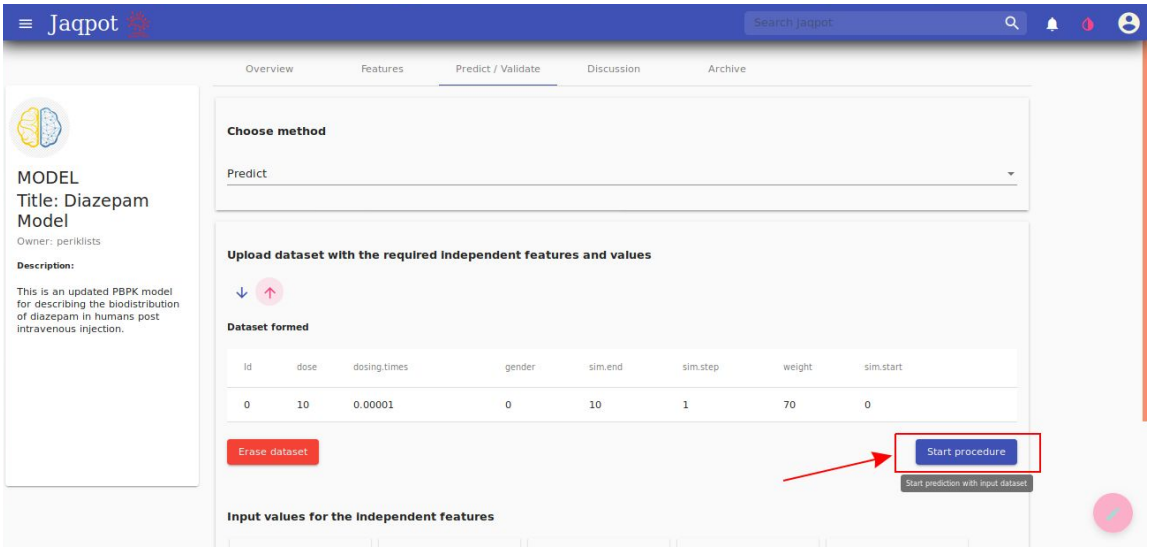



**Figure A10**. The upload process unlocks the 'start procedure' button.

As it is clear in Figure A10, once the csv is uploaded the user can review the filled values and then press the 'start procedure' button to initiate the prediction, or click on 'Erase dataset' if a mistake is spotted.

It has to be noted that if a model supports vectorized input (e.g. a vector of multiple doses), the user can only provide this kind of input only through the GUI in the following format: [value1, value2, ...] (Figure A11). In this case, the 'Start' button on the bottom right end of the screen appears only after all values have been filled in, so NULL values are not feasible.

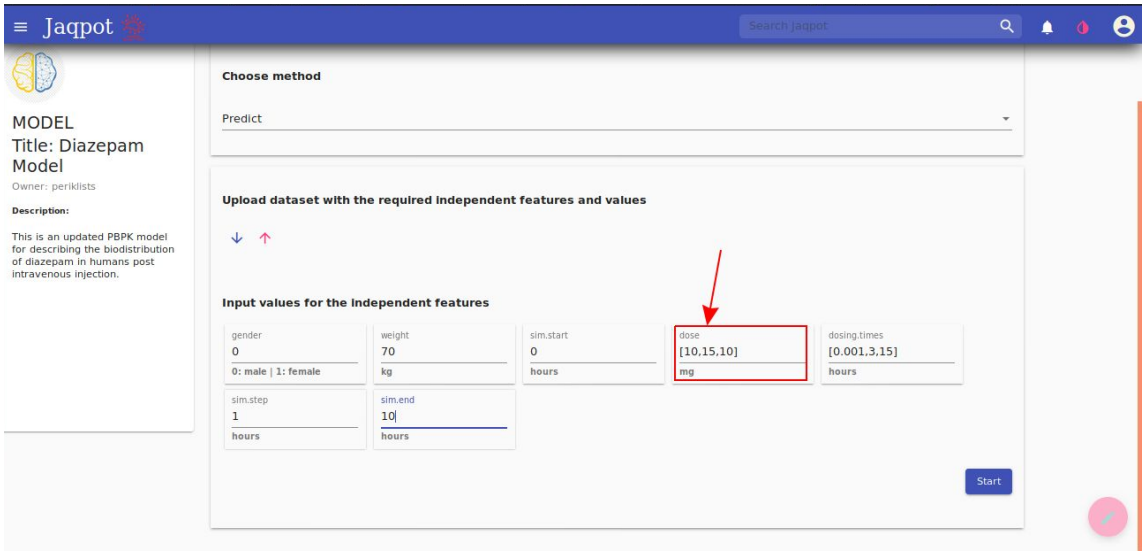



**Figure A11**. Example of a vector input on the GUI.

When the prediction process is initiated, a small log is generated on the screen and, if no error occurs, the user can proceed to the results by clicking on the double arrow icon on the bottom of the screen (Figure A12).

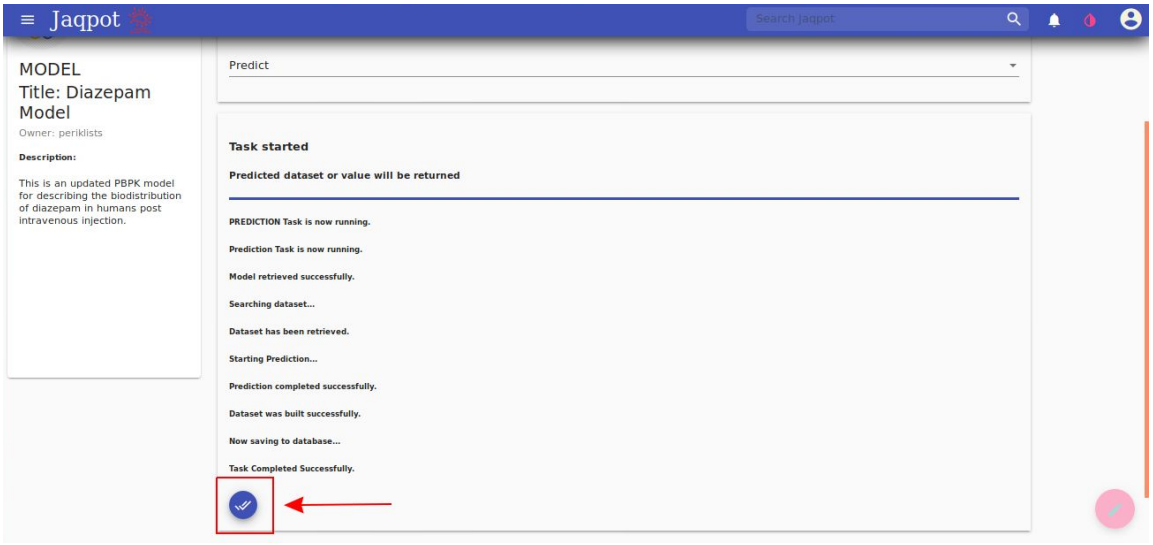



**Figure 12**. Click on the 'View prediction' icon to obtain the predictions.

The results are given on a tabular format on the GUI and can be downloaded for further processing by clicking on the download button (Figure A13).

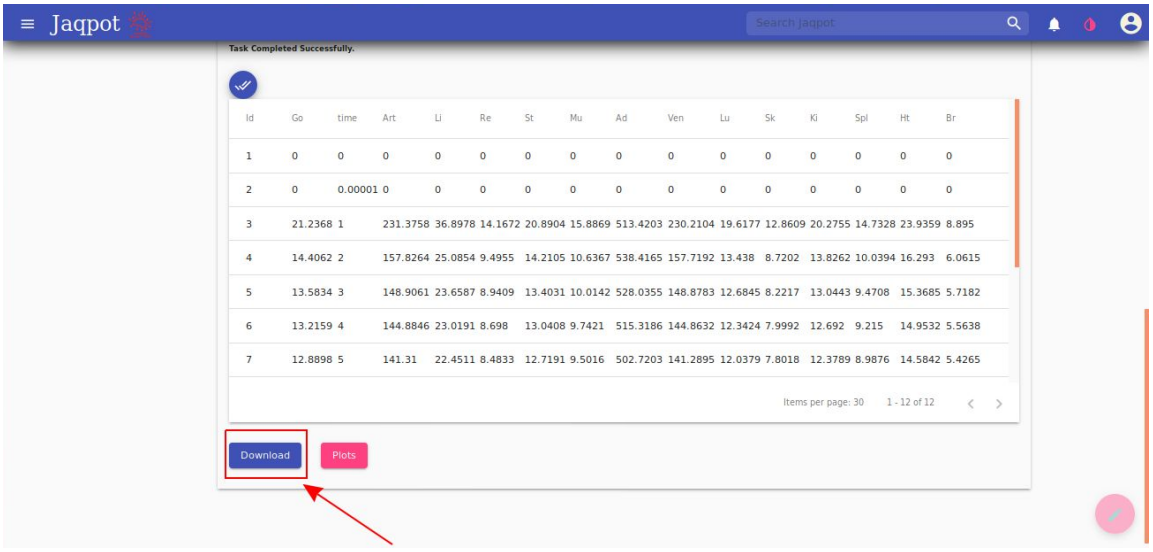



**Figure A13**. The 'Download' button allows downloading the results in a csv format.

The 'Plots' button which is positioned right next to the 'Download' button allows the user to produce plots by selecting the desired dependent features using the drag-and-drop technique (Figure A14). The desired plot then appears under the predictions (Figure A15).

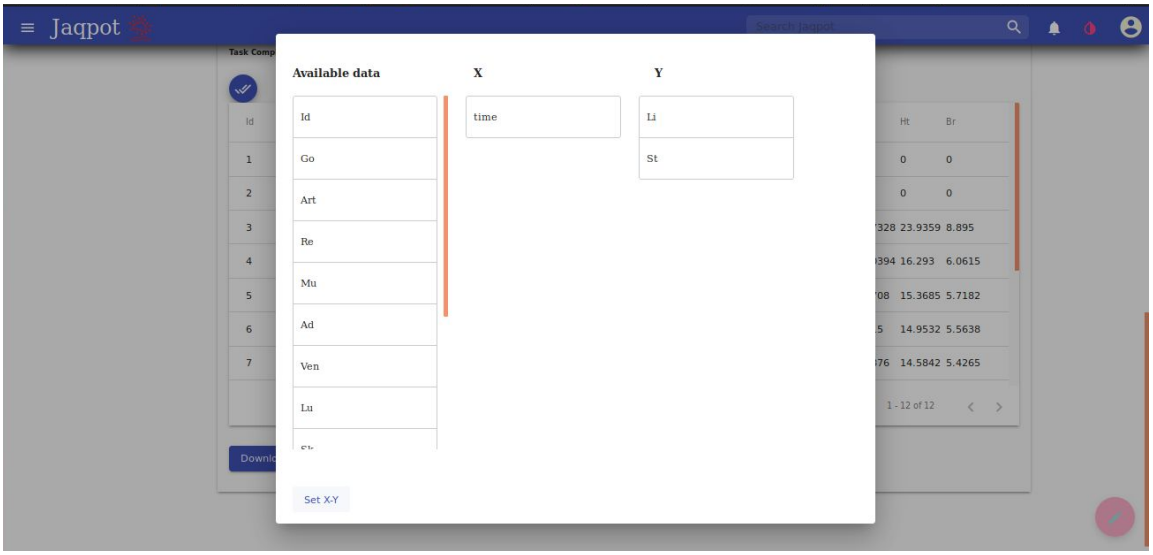



**Figure A14**. A plot can be generated by dragging and dropping the desired dependent features on the x and y-axis respectively.

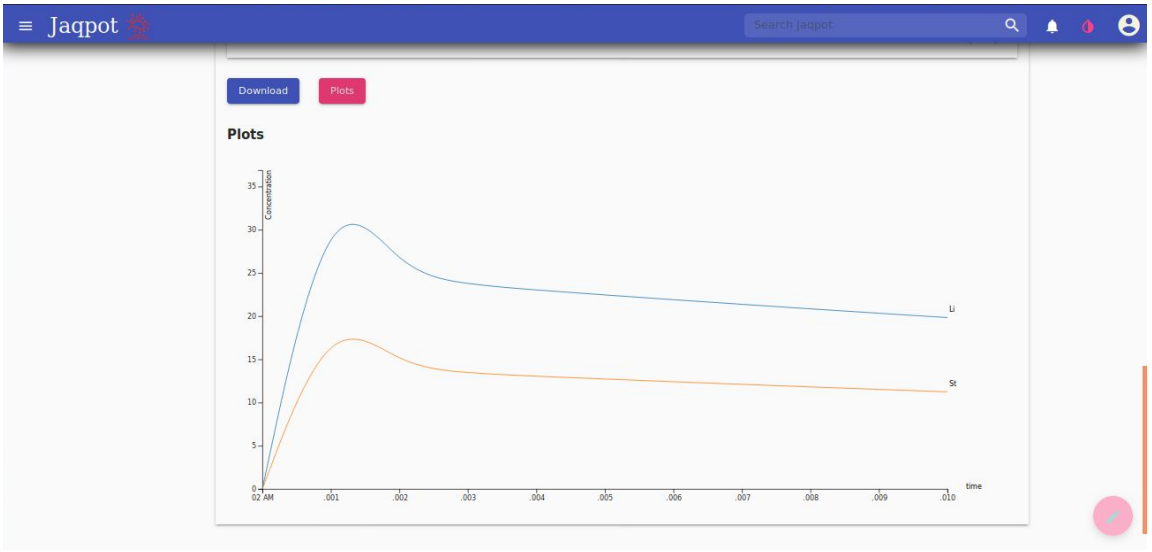



**Figure A15**. Plot that shows the concentration of diazepam in the liver and stomach compartment.

Finally, the user can add comments and remarks or ask a question regarding the model under the 'Discussion' tab (Figure A16).

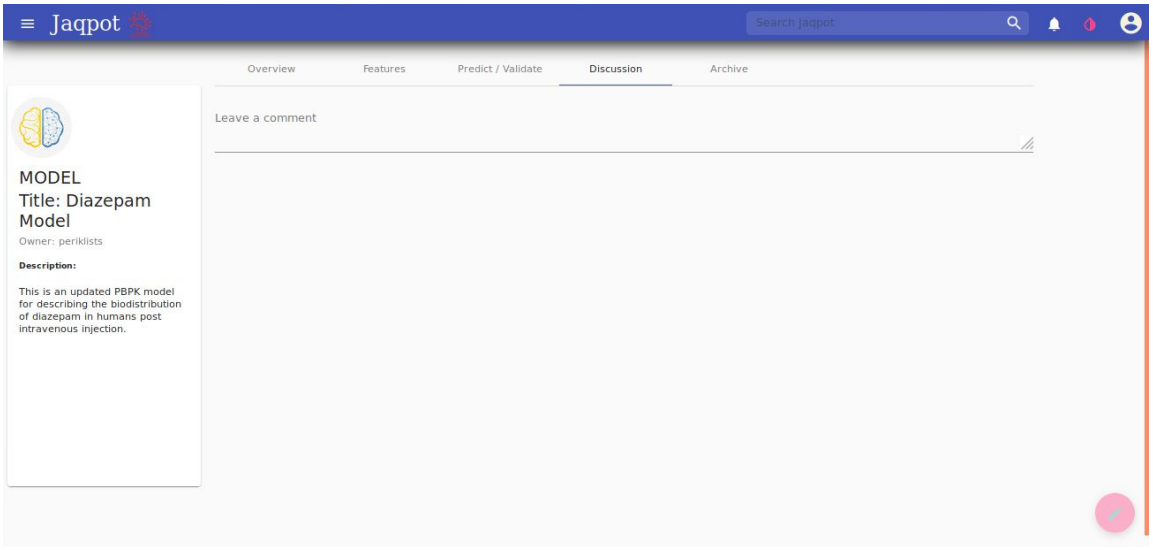



**Figure A16**. 'Discussion' tab of the Diazepam model.