

OpenRiskNet

RISK ASSESSMENT E-INFRASTRUCTURE

Case Study

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

SUMMARY	2
DESCRIPTION	4
Implementation team	4
Case Study objective	4
DEVELOPMENT	5
Databases and tools	5
Service integration	5
Technical implementation	6
OUTCOMES	8
Data access and curation workflow using Squonk	8
Data curation for the LTKB dataset	9
Merging LTKB data with TG-GATEs	9
Combining of data and knowledge sources	9
Finding similar data-rich compounds for read across	14
Data mining workflow for carcinogenicity predictions	16
DataCure webinar	18
REFERENCES	19

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.

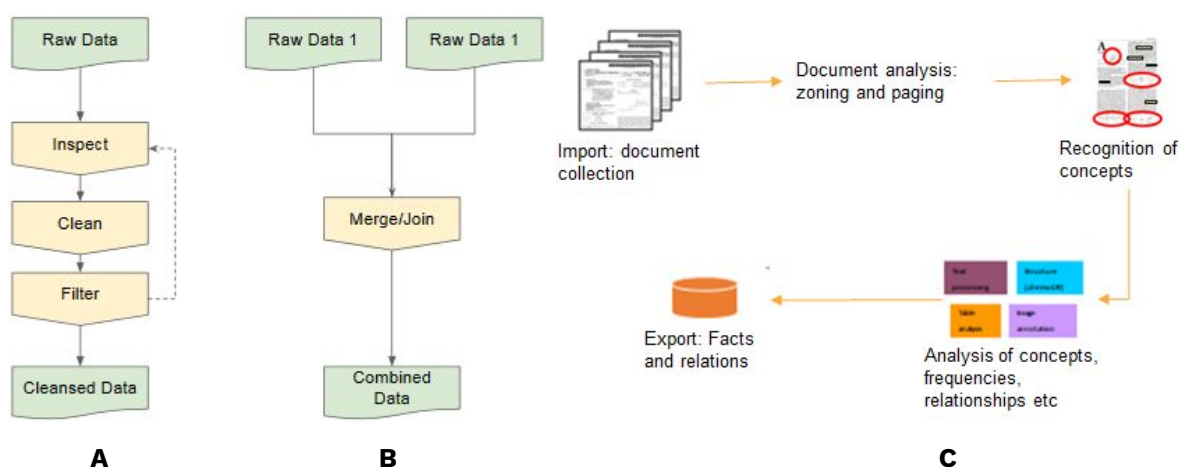


Figure 1. A: Raw data curation workflows; B: Data merging workflow C: Text mining workflow in general

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: [SCAView 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 SCAView, 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 a 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. Fagnet 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.

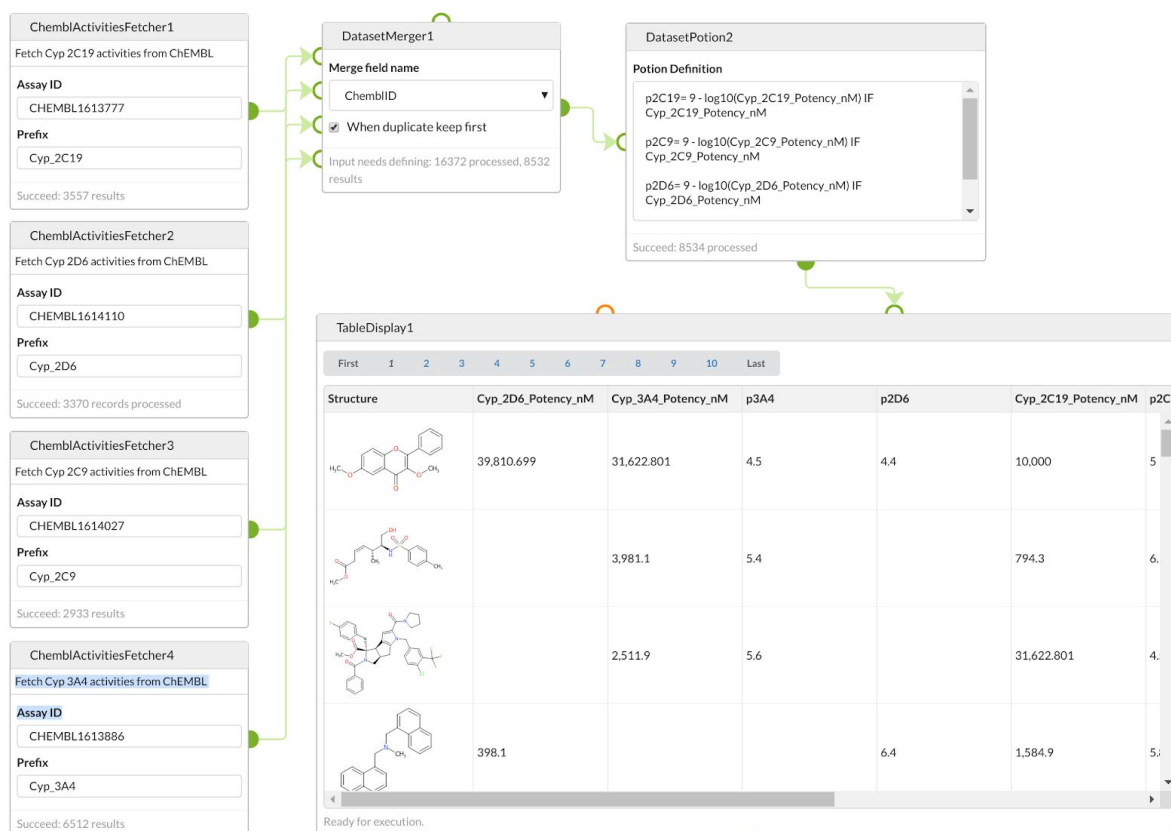


Figure 1. Squonk notebook to curate the P450 datasets available at <https://github.com/OpenRiskNet/notebooks/tree/master/DataCure/P450>

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.

registry.prod.openrisknet.org/swaggerui?service=https%3A%2F%2Fapi.staging.kit.cloud.douglasconnect.com%2Fdatasets%2F97f64415-8c1b-426d-ac06...

| Design Type | Format | Organism | Tissue | Cell Name | |---|---|---|---| | growth reporter - real-time cell-growth kinetics | cell-based - cell-based format | human | breast | T47D |

| Biological Process | Target Family | Target Type | |---|---|---|---| | cell proliferation | nuclear receptor - steroidal | pathway - pathway-specified |

References

209. Xing JZ, Zhu L, Gabos S, Xie L. Microelectronic cell sensor assay for detection of cytotoxicity and prediction of acute toxicity. *Toxicol In Vitro*. 2006 Sep;20(6):995-1004. Epub 2006 Feb. PubMed PMID: [16481145](#).

210. Rotroff DM, Dix DJ, Houck KA, Kavlock RJ, Knudsen TB, Martin MT, Reif DM, Richard AM, Sipes NS, Abassi YA, Jin C, Stampfl M, Judson RS. Real-time growth kinetics measuring hormone mimicry for ToxCast chemicals in T-47D human ductal carcinoma cells. *Chem Res Toxicol*. 2013 Jul 15;26(7):1097-107. doi:10.1021/tx400117y. Epub 2013 Jun 10. PubMed PMID: [23682706](#).

Servers

default

GET /datasets/97f64415-8c1b-426d-ac06-9a048bdc8329/versions/1/data Returns the data along with aggregation/faceting information and a total count

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	CCCC(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	HEMJKBWTPKOJG-UHFFFAOYSA-N	25812-30-0	NaN
964	ATG_PPArA_TRANS_up	DTXSID0020652	Gemfibrozil	HEMJKBWTPKOJG-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	PValue	adj.PVal	
70	gemfibrozil	HEMJKBWTPKOJG-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	HEMJKBWTPKOJG-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	HEMJKBWTPKOJG-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	HEMJKBWTPKOJG-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	HEMJKBWTPKOJG-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	HEMJKBWTPKOJG-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	HEMJKBWTPKOJG-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	HEMJKBWTPKOJG-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	HEMJKBWTPKOJG-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	HEMJKBWTPKOJG-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	HEMJKBWTPKOJG-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	HEMJKBWTPKOJG-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	HEMJKBWTPKOJG-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

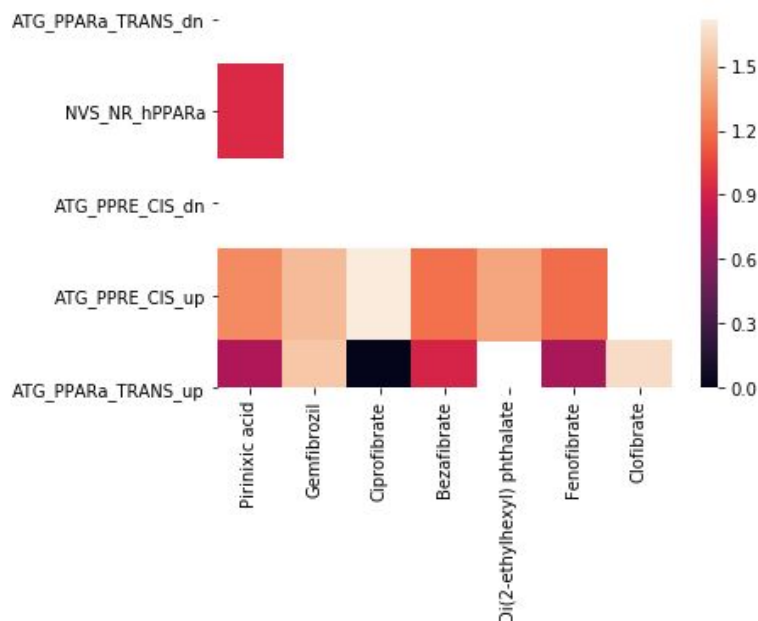


Figure 8. Heatmap of IC50 values extracted from ToxCast for the stressors of AOP 37

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, Perhydropridine, Piperidin [German], Piperidine [UN2401] [Corrosive], Pyridine, hexahydro- UN2401, 80645_FLUKA, Azacyclohexane, C01746, Hexahydropridine, 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-m/s1	NQCQWIFBJYNOQM-SSDOTTWSA-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/c6-4-2-7-3-5/10/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'])

```

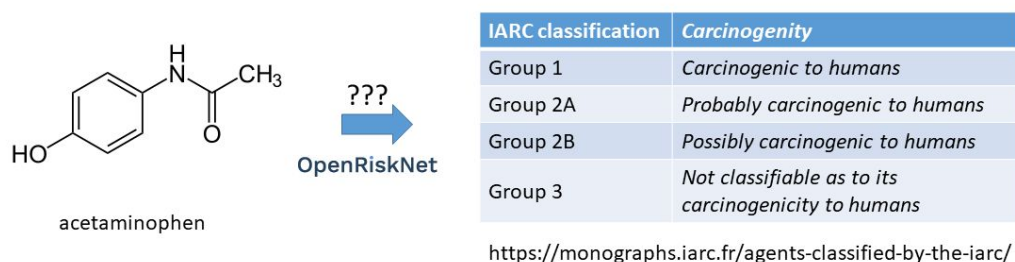
	Assay	DTXSID	Substance name	InChI key	CAS	IC50
5016	TOX21_AhR_LUC_Agonist	DTXSID6021165	Piperidine	NQRYJNQNLNOLGT-UHFFFAOYSA-N	110-89-4	None

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).

DataCure – Data curation and creation of pre-reasoned datasets and searching ¹⁾



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 ([SCAView](#))
4. further analyze documents with NLP (SCAView -> UIMA) in order to find evidence sentences supporting the classification of a compound

Text Mining workflow:

Task:

- Identify the concept of acetaminophen (definition, identifiers, synonyms)
- Find all relevant documents in the context of acetaminophen and carcinogenicity
- What are the most relevant statements

Technology:

- Semantic index of PubMed/PMC (> 20 terminologies)
- Solr index + OLS index + UIMA pipeline

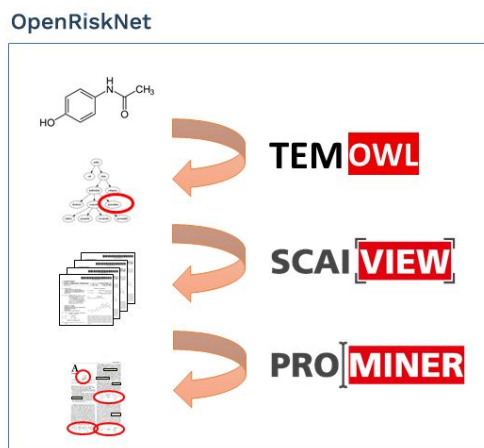


Figure 12. Problem description: into which class belongs acetaminophen?

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](https://pubchem.ncbi.nlm.nih.gov/compound/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 file1].',

'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 2018¹. 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.