Organisation for Economic Co-operation and Development

ENV/JM/MONO(2020)20/ANN3

Unclassified English - Or. English

24 September 2020

ENVIRONMENT DIRECTORATE
JOINT MEETING OF THE CHEMICALS COMMITTEE AND THE WORKING PARTY ON CHEMICALS, PESTICIDES AND BIOTECHNOLOGY

ANNEX III: DETAILED DESCRIPTION OF IN SILICO MODELS WITHIN CASE STUDY ON THE USE OF INTEGRATED APPROACHES TO TESTING AND ASSESSMENT FOR PREDICTION OF A 90 DAY REPEATED DOSE TOXICITY STUDY (OECD 408) FOR 2-ETHYLBUTYRIC ACID USING A READ-ACROSS APPROACH FROM OTHER BRANCHED CARBOXYLIC ACIDS

Series on Testing and Assessment
No. 324

The corresponding monograph to this annex is available under the following cotes: ENV/JM/MONO(2020)20.

JT03465707

Table of Contents

1. Assessment of structural similarity

Structural similarity was calculated by using a workflow integrated into a KNIME analytic platform (https://www.knime.com/knime-software). We used smiles codes as input parameter.

Smiles codes were extracted from high quality open source databases. In this case, the majority comes from the US EPA chemical dashboard (https://comptox.epa.gov/dashboard) and one from CHEMID Plus (https://chem.nlm.nih.gov/chemidplus/, Table 1). Prior to the calculation of structural fingerprints, the smiles codes for all analogues were quality controlled, a correction was not needed.

We used a fingerprint method, which involves encoding the structural information within a molecule as a bit string in which each bit indicates the presence of ("1") or absence ("0") of a particular molecular feature. Molecular fingerprints per compound were calculated with RDKIT using a MACCS keys (Molecular ACCess System). MACCS keys (Durant JL *et al.* 2002) are a classical fingerprint in chemoinformatics consisting of a dictionary of 166 structural fragments, as well as UNITY fingerprints (Patterson DE *et al.*) that assemble atom pathways of predefined lengths.

In order to assess structural similarity between compounds we used the Dice coefficient (Gillet *et al.* 2003, Dice 1945). The Dice coefficient (S) calculates similarity comparing pairs of compounds, e.g. A and B (Equation 1). It uses three values, the number of bits set to 1 in both fingerprints (value a and b) and the number of shared bits set to 1 between the two compounds (value c, equation 1). Other algorithms like Tanimoto did result in similar values.

Equation 1: $S(A,B) = \frac{2c}{a+b}$

References:

Gillet VJ and Leach AR (2003) An Introduction to Chemoinfromatics. Kluwer Academic Publishers Drdrecht, The Netherlands, 103.

Dice, Lee R. (1945). "Measures of the Amount of Ecologic Association Between Species". *Ecology.* 26 (3): 297–302. doi:10.2307/1932409

Durant JL, Leland BA, Henry DR, *et al.* : Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci.* 2002;42(6):1273–1280. 10.1021/ci010132r

Patterson DE, Cramer RD, Ferguson AM, *et al.* (1996) Neighborhood behavior: a useful concept for validation of "molecular diversity" descriptors. *J Med Chem.* 39(16):3049–3059. 10.1021/jm960290n

2. Profiling modules from the OECD toolbox

GENERAL REMARKS:

Profilers from the OECD QSAR Toolbox can be used during the formation of chemical categories in order to group together chemicals on a mechanistic basis (e.g., chemicals following the same mode of toxicological action). Profilers flag the "presence of alerts" and the absence of known alerts should not be interpreted as a lack of toxicity. The profilers from the OECD QSAR Toolbox are not documented by QSAR Model Reporting Formats (QMRF) since their output must not be interpreted as a QSAR prediction but only as information upon which a robust chemical category can be formed or pertinent chemical analogues can be identified. The results presented in this submission were obtained thanks to the OECD QSAR Toolbox v4.2.The documentation on the profilers is reported below

DNA BINDING BY OASIS:

All the chemicals of interest do not fire a known alert. The documentation associated with the profiler reports only information about true positives and false positives. According to this piece of information the positive predictive value is equal to 56%.

About section of a profiler
Name of the profiler
DNA binding by OASIS
Developer; Donator; date; version
Developer: Laboratory of Mathematical Chemistry (LMC), Bourgas, Bulgaria ***Donator:*** Laboratory of Mathematical Chemistry (LMC), Bourgas, Bulgaria ***Version:*** 1.5 December 2017
Relevance/Applicability to endpoint(s)
DNA binding by OASIS is a general mechanistic profiler, consisting of list of structural alerts. The profiler relies on the alerts from the AMES Ames mutagenicity TIMES model. The profiler is based on 85 structural alerts responsible for interaction of chemicals with DNA. The scope of the profiler is to investigate presence of alerts within target molecules which may interact with DNA. The list of 85 structural alerts has been separated into eight mechanistic domains. The profiling result assigns a target to the corresponding structural alert, mechanistic alerts and domain.
Relevance/Applicability to particular chemical classes
This profiler is applicable to those organic chemicals that have presence of at least one of the 85 DNA binding alerts specified within the profiler. The presence of DNA binding alerts is not bounded with parametric ranges; it is based on structural boundaries only.
Approach used to develop the profiler - Concise but informative description of:
a) The overall rationale: The aim of the profiler is to investigate presence of alerts within target molecules which may be responsible for interaction with DNA.

b) The criteria or the method applied for analysing the training set/the pool of chemicals that inform the profiler: The trainng set chemicals have been analysed based on identification for presence or absence of DNA binding alert.

c) Source of the data/knowledge and total number of chemicals included in the analysis: The profiler was based on the 85 structural alerts responsible for interaction with DNA analysed in Ames Mutagenicity model. The list of 85 structural alerts has been separated into eight mechanistic domains.

d) Literature references:

1. Mekenyan, O., Dimitrov, S., Serafimova, R., Thompson, E., Kotov, S., Dimitrova, N., and Walker, J. (2004) Identification of the structural requirements for mutagenicity by incorporating molecular flexibility and metabolic activation of chemicals I: TA100. Chem. Res. Toxicol. 17, 753-766.

2. Serafimova, R., Todorov, M., Pavlov, T., Kotov, S., Jacob, E., Aptula, A., and Mekenyan, O. (2007) Identification of the structural requirementsfor mutagencitiy, by incorporating molecular flexibility and metabolic activation of chemicals. II. General Ames mutagenicity model. Chem. Res. Toxicol. 20, 662-676.

Summary description of profiles/alerts within the profiler

Profile/structural alert	Number of analysed chemicals	Number of chemicals associated with Ames mutagenicity
Nitro Azoarenes and p-Substituted Azobenzenes	74	46/74
Nitrobiphenyls and Bridged Nitrobiphenyls	38	28/38
Conjugated Nitroalkenes and Five-Membered Aromatic Nitroheterocyclics	36	36/36
Nitroaniline Derivatives	48	40/48
Fused-Ring Nitroaromatics	46	42/46
Nitroarenes with Other Active Groups	40	26/40
Nitroalkanes	5	4/5
Nitrophenols, Nitrophenyl Ethers and Nitrobenzoic Acids	45	24/45
p-Substituted Mononitrobenzenes	13	11/13
Polynitroarenes	30	24/30
N-Aryl-N-Acetoxy(Benzoyloxy) Acetamides	6	5/6
Amino Anthraquinones	27	10/27
Fused-Ring Primary Aromatic Amines	73	25/73
p-Aminobiphenyl Analogs	32	15/32
Single-Ring Substituted Primary Aromatic Amines	63	13/63
Hydrazine Derivatives	54	38/54

Alfa,Beta-Unsaturated aldehydes	31	10/31
Specific Acetate Esters	68	13/68
Alkylphosphates, Alkylthiophosphates and Alkylphosphonates	46	13/46
Diazenes and Azoxyalkanes	3	2/3
Arenediazonium Salts	2	2/2
Organic Peroxy Compounds	44	28/44
Sulfonyl Halides	8	4/8
Thiols	34	7/34
N-Acetoxyamines	26	25/26
Alkylnitrites	8	7/8
Diazoalkanes	8	8/8
Quinoneimines	11	4/11
Polarized Haloalkene Derivatives	13	10/13
Haloisothiazolinones	1	1/1
Haloalkane Derivatives with Labile Halogen	37	28/37
Sultones	1	1/1
Vicinal Dihaloalkanes	41	15/41
Acyl Halides	26	6/26
Monohaloalkanes	8	7/8
Haloalkane Derivatives Containing Chain Heteroatom	76	58/76
Haloalkene Derivatives with Electron-Withdrawing Groups	15	9/15
Geminal Polyhaloalkane Derivatives	100	45/100
Alpha-Haloethers	8	6/8
Specific Imine and Thione Derivatives	24	9/24
Dicarbonyl Compounds	13	10/13
Quinoline Derivatives	51	13/51
Sulfonyl Azides	1	1/1
Pyrrolizidine Derivatives	5	2/5
Aminoacridine DNA Intercalators	29	27/29
Epoxides and Aziridines	75	52/75
Quinones and Trihydroxybenzenes	122	54/122
Four- and Five-Membered Lactones	28	8/28
C-Nitroso Compounds	12	8/12
N-Nitroso Compounds	48	29/48
Sulfonates and Sulfates	33	28/33
N-Acyloxy(Alkoxy) Arenamides	30	30/30
Haloalcohols	15	13/15

Acyclic Triazenes	19	12/19
Nitrogen and Sulfur Mustards	47	44/47
Polycyclic Aromatic Hydrocarbon and Naphthaleneimide Derivatives	47	15/47
Coumarins	30	8/30
N-Hydroxylamines	65	37/65
DNA Intercalators with Carboxamide and Aminoalkylamine Side Chain	118	39/118
Halofuranones	19	19/19
Anthrones	6	4/6
Triarylimidazole and Structurally Related DNA Intercalators	9	8/9
Hydroxamic Acids	6	5/6
Haloalkene Cysteine S-Conjugates	7	7/7
Acridone, Thioxanthone, Xanthone and Phenazine Derivatives	27	24/27
Flavonoids	6	3/6
N,N-Dialkyldithiocarbamate Derivatives	8	7/8
Similar to other profilers		
This profiler is similar to the general mechanistic profiler "***DNA binding by OECD***" and endpoint specific profiler "***DNA alerts for AMES by OASIS***" and "***DNA alerts for CA and MN by OASIS***". Also, it is similar to the "***In vitro mutagenicity (Ames test) alerts by ISS***" profiler to a certain extent. However, this profiler is general mechanistic and includes structural boundaries which are defined based on chemicals functionalities which may interact with DNA from theoretical point of view. These rules could be considered as necessary conditions for eliciting positive Ames effects and they are not supported by experimental data. Hence, general mechanistic profilers could not be considered as alerts for Ames mutagenicity. For example, Single-Ring Substituted Primary Aromatic Amines in the "***DNA binding by OASIS***" profiler includes only aniline moiety to describe potential reactivity towards DNA. However, for increasing probability of chemicals to cause positive Ames mutagenicity effect, additional structural requirements not included in this profile need to be defined. Because of that, general mechanistic DNA binding profilers are considered as suitable for formation of chemical categories for read across rather than use them as SARs.		
Short description of update version		
SMARTS language for describing molecular patterns, i.e. structural boundaries, structural alerts has been implemented in OECD QSAR Toolbox 4.0. As a result, "***DNA binding by OASIS***" has been rewritten. Only small distinctions are expected in the profiling results between Toolbox v.3.4 and v.4.0 due to different interpretation of the molecular structures, e.g. for heterocyclic/heteroaromatic compounds. Further general modifications are as follows: 1. Quinone Methides - modified - presence of H-atom at beta position towards carbonyl atom 2. Alkylnitrites - modified - prohibition (expressed with NOT) for nitro group and introduction of enumeration containing H-atom and C{sp3} atom		

3. Conjugated Nitroalkenes and Five-Membered Aromatic Nitroheterocyclics – new name – two alerts are united in one alert – former Conjugated Nitro Compounds and Five-Membered Aromatic Nitroheterocycles 4. Nitro Azoarenes and p-Substituted Azobenzenes – the category has a new name and new query for p-Substituted Azobenzenes is added 5. Quinoxaline-Type 1,4-Dioxides - modified – a mask forbidding fused aromatics is added The modification for QSAR Toolbox v.4.2 is: 1. The mechanistic justification for the category Sulfonul halides is replaced with the correct one. 2. The category Haloalkane Derivatives with Labile Halogen is slightly modified in order a cyclic double bond to meet the criteria too.
Disclaimer
The structural boundaries used to define the chemical classes (e.g. "Alcohol" – chemical class from "Organic functional group" profiler) or alerting groups responsible for the binding with biological macromolecules (e.g. "Aldehydes" – structural alert for protein binding), represent structural functionalities in the molecule which could be used for building chemical categories for subsequent data gap filling. They are not recommended to be used directly for prediction purposes (as SARs).

DNA BINDING BY OECD:

All the chemicals of interest do not fire a known alert. The documentation associated with the profiler does not report information on predictivity. According to the comparative work by Cassano *et al.* (2014) if the absence of alert is interpreted as a negative prediction, this profiler has a sensitivity comprised between 63% and 65% and a Matthew's correlation coefficient comprised between 0.20 and 0.28. This performance is far from being optimal but it must be noted that the other evaluated models were characterised by a comparable performance.

Reference:

Cassano *et al.*, (2014) J Environ Sci Health C Environ Carcinog Ecotoxicol Rev. 2014;32(3):273-98. doi: 10.1080/10590501.2014.938955.

About section of a profiler
Name of the profiler
DNA binding by OECD
Developer; Donator; date; version
Developer: School of Pharmacy and Chemistry, Liverpool John Moores University, UK ***Donator:*** European Chemicals Agency (ECHA); Organisation for Economic Co-operation and Development (OECD) ***Version:*** 2.3 December 2016
Relevance/Applicability to endpoint(s)

This profiler is intended to be used for the assessment of endpoints in which covalent binding to DNA has been shown to be the molecular initiating event for low molecular weight chemicals. The profiler has been developed from mechanistic knowledge of the electrophilic chemistry of covalent DNA binding – importantly it has been developed from a systematic review of the literature and not from the analysis of a single toxicological dataset.
Relevance/Applicability to particular chemical classes
This profiler is applicable only to organic chemicals that have a molecular weight less than 1000 g/mol. It is applicable only to the chemical classes for which it contains structural alerts; the absence of a structural alert should not be taken as an absence of toxicity.
Approach used to develop the profiler - Concise but informative description of:
a) The aim of the profiler was to identify structural alerts associated with organic, low molecular weight chemicals capable of forming covalent bonds with DNA. The structural alerts were derived from knowledge of the molecular initiating event - covalently binding to DNA. It was developed from a systematic review of the literature, rather than from the analysis of a single toxicological dataset.
b) The profiler was developed from a mechanistic rationale that the molecular initiating event for covalent bond formation with DNA. Importantly, this was achieved by reviewing the literature relating to the chemistry, rather than an analysis of toxicological datasets.
c) The profiler was developed from an extensive review of the literature relating to the chemistry of covalent bond formation with DNA. A full list of the literature included can be found in the reference listed in section d.
d) An overview of the mechanistic chemistry and underlying principles of the structural alerts within this profiler can be found in: Enoch *et al* (2010) *A review of the electrophilic reaction chemistry involved in covalent DNA binding.* Critical Reviews in Toxicology, 40, p728-748
Summary description of profiles/alerts within the profiler
It is not possible to provide metrics relating to this profiler as it was not developed from an analysis of toxicological datasets. It was developed from an extensive review of the chemistry related to the formation of a covalent bond between a low molecular weight chemical and DNA.
Similar to other profilers
A number of related endpoint specific profilers exist in the OECD QSAR Toolbox relating to genotoxicity. The ***DNA binding by OECD*** profiler should be used first, with endpoint specific profilers (which have been developed from an analysis of toxicological data) being used to sub-categorise, where possible.
Short description of update version
SMARTS language for describing molecular patterns, i.e. structural boundaries, structural alerts has been implemented in OECD QSAR Toolbox 4.0. As a result ***DNA binding by OECD*** profiler has been rewritten but without modifying the knowledge it is based on. Distinctions are expected in the profiling results between Toolbox v.3.4 and v 4.0 due to different interpretation of the molecular structures, e.g. for heterocyclic/heteroaromatic compounds and the new 2D redactor which allows to define the structure boundaries more correctly according to the description of the categories. An example for category with possible inconsistencies between TB 3.4 and TB 4.0 is the Aromatic azo category. The profiling results in TB 4.0 are expected to be more accurate than these of TB 3.4.

Disclaimer
The structural boundaries used to define the chemical classes (e.g. "Alcohol" – chemical class from "Organic functional group" profiler) or alerting groups responsible for the binding with biological macromolecules (e.g. "Aldehydes" – structural alert for protein binding), represent structural functionalities in the molecule which could be used for building chemical categories for subsequent data gap filling. They are not recommended to be used directly for prediction purposes (as SARs).

EYE IRRITATION/CORROSION EXCLUSION RULES by BfR:

The OECD QSAR Toolbox does not provide any documentation for this profiler. The exclusion rules for eye irritation/corrosion were evaluated by Tsakovska *et al.* (2005) and 87% of inoffensive chemicals are recognised as such.

Reference:

Tsakovska *et al.* (2005) Evaluation of (Q)SARs for the prediction of Eye Irritation/Corrosion Potential Physicochemical exclusion rules (EUR 21897 EN)

https://eurl-ecvam.jrc.ec.europa.eu/laboratories-research/predictive_toxicology/doc/Evaluation_of_Eye_Irritation_QSARs.pdf

EYE IRRITATION/CORROSION INCLUSION RULES by BfR:

The OECD QSAR Toolbox does not provide any documentation for this profiler. Only "2-Ethyl-1-hexanol" meets a rule but this chemical is not within the remit of the read-across mock submission. Tsakovska tried in 2007 to evaluate these rules but there were too few available chemicals for a robust assessment of predictivity.

Reference:

Tsakovska *et al.* (2007) Evaluation of SARs for the prediction of eye irritation/corrosion potential: structural inclusion rules in the BfR decision support system. SAR QSAR Environ Res. 2007 May-Jun;18(3-4):221-35.

SKIN IRRITATION/CORROSION EXCLUSION RULES by BfR:

The OECD QSAR Toolbox does not provide any documentation for this profiler. Only "2 –Propylheptanoic acid" flags an alert based on estimated melting point and the prediction of this property is widely known to be particularly uncertain (Dearden *et al.*, 2013). The OECD QSAR Toolbox does not provide any documentation for this profiler. Nevertheless, the exclusion rules for skin irritation/corrosion were evaluated by Rorije *et al.* (2005) and only 2% of the predictions were false negatives.

References:

Dearden *et al.* (2013). QSPR prediction of physico-chemical properties for REACH. SAR QSAR Environ Res. 2013;24(4):279-318. doi: 10.1080/1062936X.2013.773372.

Rorije and Hulzebos(2005) Evaluation of (Q)SARs for the prediction of Skin Irritation/Corrosion Potential–Physico-chemical exclusion rules - https://eurl-

ecvam.jrc.ec.europa.eu/laboratoriesresearch/predictive_toxicology/information-sources/qsar-documentarea/Evaluation_of_Skin_Irritation_QSARs.pdf

SKIN IRRITATION/CORROSION INCLUSION RULES by BfR:

The OECD QSAR Toolbox does not provide any documentation for this profiler. Only "2-Ethyl-1-hexanol", "4-Pentenoic acid 2-fluoro-2-propyl" and "2-Propylheptanoic acid" do not meet any inclusion rule. Only the last chemical is of interest for the mocking submission. The remaining chemicals are identified as obviously identified as "aliphatic acids". Gallegos -Saliner *et al.*, (2007) evaluated these rules and estimated a Positive Predictive Value of 68.2% for skin irritation and 94.7% for skin corrosion

Reference:

Gallegos-Saliner *et al.* (2007). Evaluation of SARs for the prediction of skin irritation/corrosion rus in the BfR decision support system. SAR QSAR Environ Res. 2007 May-Jun;18(3-4):331-42.

PROTEIN BINDING POTENCY h-CLAT:

All the chemicals of interest do not fire a known alert. According to the profiler's documentation the positive predictivity of the alerts ranges from 0.5 to 1 and the majority if the alerts are characterised by the highest possible positive predictivity. Nevertheless, in some cases, the evaluation was carried out in the presence of few chemicals and the robustness of this evaluation is questionable.

About section of a profiler
Name of the profiler
Protein Binding Potency h-CLAT
Developer; Donator; date; version
Developer: Laboratory of Mathematical Chemistry (LMC), Bourgas, Bulgaria ***Donator:*** Laboratory of Mathematical Chemistry (LMC), Bourgas, Bulgaria ***Version:*** 1.0 December 2016
Relevance/Applicability to endpoint(s)
This profile is built in relation with the implementation of the adverse outcome pathway (AOP) for skin sensitisation. It is developed on the base of data derived from the human cell line activation (h-CLAT) assay (). The h-CLAT is an *in vitro* method proposed to address the third key event (dendritic cell activation) of the skins sensitisation AOP by quantifying changes in the expression of cell surface markers associated with the process of activation of DC (i.e. CD86 and CD54), in the human leukemia cell line THP-1, following exposure to sensitisers (8). The measured expression levels of CD86 and CD54 cell surface markers are then used for supporting the discrimination between skin sensitisers and non-sensitisers.

The profile contains 30 structural alerts extracted from 223 chemicals with positive/negative data values. In cases of chemicals with available more than one value, the positive value has been used in the boundaries development.
The profiling results outcome assigns a target to the corresponding potency category based on matched structural criteria.

Relevance/Applicability to particular chemical classes

This profiler is applicable to chemicals containing at least one alert listed in the profiler. Absence of alerts in the molecular structure may be associated to inability of chemicals to interact with proteins, or may be due to a lack of mechanistic knowledge. Therefore, the 'No structural alert' flag is not equivalent to a negative prediction.

Approach used to develop the profiler - Concise but informative description of:

The presented list of structural requirements has been extracted from the 163 chemical with experimental data produced by the h-CLAT test.

Literature references:

- Ashikaga, T., Sakaguchi, H., Sono, S., Kosaka, N., Ishikawa, M., Nukada, Y., Miyazawa, M., Ito, Y., Nishiyama, N. and Itagaki, H. 2010. A comparative evaluation of *in vitro* skin sensitization tests: the human cell-line activation test (h-CLAT) versus the local lymph node assay (LLNA). Altern. Lab. Anim. 38:275-84.
- Sakaguchi, H., Ashikaga, T., Miyazawa, M., Kosaka, N., Ito, Y., Yoneyama, K., Sono, S., Itagaki, H., Toyoda, H. and Suzuki, H. 2009. The relationship between CD86/CD54 expression and THP-1 cell viability in an *in vitro* skin sensitization test-human cell line activation test (h-CLAT). Cell Biol. Toxicol. 25: 109-126.

Summary description of profiles/alerts within the profiler

Positive predictivities shown in the table below are calculated by applying the profiler to the database "Dendritic cells COLIPA" (QSAR Toolbox 4.0). This database contains results of h-CLAT test (CD86/CD54 expression) on over 220 chemicals (around 110 positive, 50 negatives and the rest chemicals (about 70) are with Not determined data).
In the analysis, a chemical is defined as 'positive', if at least one alert is present in its molecular structure. The calculation of the predictivity of each individual alert is based on the CD86 or CD54 expression, if a chemical has more than one value worst case scenario is applied.

Profile alert	
1,2- and 1,3-Dicarbonyls	
Acid anhydrides	
Activated aryl esters	
Activated halo-benzenes	
Alkyl halides	
alpha, beta-Unsaturated aldehydes	
alpha, beta-Unsaturated esters	
alpha, beta-Unsaturated ketones	
alpha-Activated benzyls	
beta-Lactam	
Carbamates	
C-Nitroso compounds	
Cyclopropenones	

Di-substituted alpha, beta-unsaturated aldehydes	
Epoxides	
Five-membered heterocyclic urea	
Halo-substituted dinitriles	
Isothiazolinone derivatives	
Lactones	
Monocarbonyls	
Monohaloarenes	
N-Chloro-sulphone amides	
N-substituted aromatic amides	
Pyranones, Pyridones and related chemicals	
Quinones, quinone (di)imines, quinone methides and nitro quinone imines	
Sulfates	
Thiols and disulfides	
Vinyl pyridines	
Counter category: No alert found	
Similar to other profilers	
This profiler is similar to Keratinocyte gene expression profiler. The both profilers belong to the Endpoint specific category. They are developed in relation with implementation of AOP in QSAR Toolbox.	
Short description of update version	
Disclaimer	
The structural boundaries used to define the chemical classes (e.g. "Alcohol" – chemical class from "Organic functional group" profiler) or alerting groups responsible for the binding with biological macromolecules (e.g. "Aldehydes" – structural alert for protein binding), represent structural functionalities in the molecule which could be used for building chemical categories for subsequent data gap filling. They are not recommended to be used directly for prediction purposes (as SARs).	

PROTEIN BINDING ALERTS FOR CHROMOSOMAL ABERRATION BY OASIS:

All the chemicals of interest do not fire a known alert.

About section of a profiler
Name of the profiler
Protein binding alerts for Chromosomal aberration by OASIS
Developer; Donator; date; version
Developer: Laboratory of Mathematical Chemistry (LMC), Bourgas, Bulgaria ***Donator:*** Laboratory of Mathematical Chemistry (LMC), Bourgas, Bulgaria

Version: 1.4
December 2017

Relevance/Applicability to endpoint(s)

The profiler is based on 33 structural alerts accounting for interactions of chemicals with specific proteins, such as topoisomerases, cellular protein adducts, etc. The scope of this profiler is to investigate the ability of target molecules to elicit clastogenicity and/or aneugenicity. Functionalities which bring about steric (or electronic) hindrance in molecules and thus impede interactions with proteins are explicitly defined and associated with some of the alerts as "inhibition" masks.

This profiler is endpoint specific and is designed to indicate chemicals that could interact with topoisomerases, cellular protein adducts. The structural alerts in the endpoint specific profiler are focussed on chemistry associated with covalent binding to proteins associated with Chromosomal aberrations. In this respect the specificity of the profiler is codded by using a specific inhibition "masks" associated with some of structural alerts. As such, this profiler should be used not as a primary grouping method, but as a secondary method for refining the primary group of chemicals. As a result, more consistent group of chemical responsible for causing chromosomal aberration could be obtained.

Relevance/Applicability to particular chemical classes

This profiler is applicable to those organic chemicals that have presence of at least one of the 33 protein binding alerts specified within the profiler. The presence of protein binding alerts is not bounded with parametric ranges; it is rather based on structural boundaries only. Chemicals not classified by the profiler are marked with "No alert found"

Approach used to develop the profiler - Concise but informative description of:

a) The aim of this profiler is to investigate the presence of alerts within the target molecules responsible for interaction with proteins such as topoisomerases, cellular protein adducts, etc.

b) The profiler consists of 33 structural alerts. The alerts are separated into 9 mechanstic domains. Some of the structural alerts belong to more than one mechanistic domain.

c) The profiler was developed from a dataset of 1082 chemicals that have experimental data for chromosomal aberration.

Summary description of profiles/alerts within the profiler

Profiler alerts	Number of analysed chemicals	Number of chemicals associated with Chromosomal aberration (Correctly predicted chemicals)	Number of Correctly predicted positive chemicals	Number of Correctly predicted negative chemicals
Isocyanates and Diisocyanates	10	8/10	5/8	3/8
Isothiocyanates	4	4/4	4/4	-
Carboxylic Acid Amides	21	20/21	7/20	13/20
Arenecarboxylic Acid Esters	25	25/25	6/25	19/25
Arenesulphonamides	21	19/21	9/19	10/19
Carbamates	22	18/22	10/18	8/18

Carboxylic acid Anhydrides	13	13/13	3/13	10/13
Hexahydrotriazine Derivatives	2	2/2	2/2	-
alpha, beta - Unsaturated Carbonyls and Related Compounds	51	42/51	13/42	29/42
Pyrazolone and Pyrazolidine Derivatives	5	4/5	4/4	0/4
alpha, beta - Unsaturated Carboxylic Acids and Esters	37	29/37	21/29	8/29
Gallic Acid Esters	3	3/3	3/3	-
Hydroxylated phenols	13	11/13	10/11	1/11
N-Substituted Aromatic Amines	35	28/35	15/28	13/28
Quinoneimines	2	2/2	2/2	-
Substituted Anilines	79	56/79	25/56	31/56
Substituted Phenols	62	58/62	33/58	25/58
Pyrimidines and Purines	18	15/18	12/15	3/15
Ethenyl Pyridines	1	1/1	1/1	-
Propargyl Alcohol Derivatives	2	2/2	2/2	-
Bipyridilium herbicides	2	2/2	2/2	-
Heterocyclic Aromatic Amines	13	12/13	12/13	0/13
Benzoquinolines and Acridines	10	10/10	7/10	3/10
Sterically Hindered Piperidine Derivatives	5	4/5	4/4	-
N-Alkyl-N-nitrosocarbamates	15	15/15	15/15	-
Alkylated nitrosoureas and nitrosoguanidines	18	18/18	18/18	-
N-Nitrosoamine Derivatives	21	20/21	7/20	13/20
Alpha-Activated Haloalkanes	13	8/13	6/8	2/8
Cyanohydrines	2	2/2	1/2	1/2
Nitrogen Mustard	6	6/6	6/6	-
alpha, omega-Dihaloalkanes	2	2/2	2/2	-
Halogenated Vicinal Hydrocarbons	21	20/21	11/20	9/20

Dialkyl Alkylphosphonates	5	5/5	5/5	-
Total: 33 Alerts	561	486/561	283/486	203/486

Counter category: No alert found

Similar to other profilers

Short description of update version

SMARTS language for describing molecular patterns, i.e. structural boundaries, structural alerts has been implemented in OECD QSAR Toolbox 4.0. As a result, ***Protein binding alerts for Chromosomal aberrations by OASIS*** has been rewritten. Only small distinctions are expected in the profiling results between Toolbox v.3.4 and v 4.0 due to different interpretation of the molecular structures, e.g. for heteroatomic compounds.
Further general modifications are as follows:
1. Carboxylic Acid Amides – a mask is added prohibiting aryl-azo-keto fragment
2. Substituted Anilines – a mask is added prohibiting aryl-azo-keto fragment
3. Substituted Phenols – masks are added to the queries
4. Hydroxylated Phenols – masks are added to the queries
5. N-Substituted Aromatic Amines - a mask is added
6. alpha, beta-Unsaturated Carbonyls and Related Compounds – a mask is added; the query for alpha,beta-Unsaturated aldehydes is modified
7. Halogenated Vicinal Hydrocarbons – the query is modified by addition of masks, enumeration and explicit H-atoms
8. Heterocyclic aromatic amines - new query is added for triazines

Modifications implemented in OECD QSAR Toolbox 4.1 are as follows:
1. Halogenated vicinal Hydrocarbons
   The general structural definition is extended by addition of the following substituents: hydrogen, hydroxyl group, sulfur, phosphorous and carbon. The addition was necessary due to expert opinion based on newly added chemicals to the training data set.
2. Heterocyclic aromatic amines
   As a result of expansion of the training set with in house tested chemicals new structural mask – trisubstituted triazine with alkyl or arylamino groups is adeded by expert judgment.
3. Substituted phenols
   The structural definition is extended by addition of new mask - alkyl and alkoxy substituted phenols. The need of this mask was expertly specified.
4. alpha,beta-Unsaturated Carbonyls and Related Compounds
   The query for alpha, beta - Unsaturated ketones was expertly modified according to available training set representatives.

The updates in the profiler are: attached local training sets to the structural alerts available in Toolbox 4.2 consist of:
1. Addition of local training sets to the corresponding structural alerts including the following information:
2. Chemical ID (CAS, Name, SMILES)

3. Representative experimental data - in case of multiple data the worst case scenario or expert judgement is used 4. Metabolic activation (without and with S9 activation) 5. Bioassay (*in vitro* Mammalian Chromosome Aberration Test) 6. References
Disclaimer
The structural boundaries used to define the chemical classes (e.g. "Alcohol" – chemical class from "Organic functional group" profiler) or alerting groups responsible for the binding with biological macromolecules (e.g. "Aldehydes" – structural alert for protein binding), represent structural functionalities in the molecule which could be used for building chemical categories for subsequent data gap filling. They are not recommended to be used directly for prediction purposes (as SARs).

PROTEIN BINDING ALERTS FOR SKIN SENSITISATION ACCORDING TO GHS:

All the chemicals of interest do not fire a known alert. The documentation does not provide an insight into the mispredictions but it provides only information on true positives and false positives. According to these statistics the Positive predictive value of the module = TP/(TP+FP) = 80% .

About section of a profiler
Name of the profiler
Protein binding alerts for skin sensitization according to GHS
Developer; Donator; date; version
Developer: Laboratory of Mathematical Chemistry (LMC), Bourgas, Bulgaria ***Donator:*** Laboratory of Mathematical Chemistry (LMC), Bourgas, Bulgaria ***Version:*** 1.1 December 2017
Relevance/Applicability to endpoint(s)
The profiler is based on recently developed OASIS TIMES model for predicting skin sensitisation according to GHS criteria. The protein binding alerts are extracted from 517 chemicals used as a training set for the model. In the current version of the profiler are available 135 alerts, classifying chemicals into GHS categories 1A and 1B. The borders of Category 1B are slightly modified in the model and respectively in the profiler. As a result the classification thresholds are the following: • Category 1A – EC3 (LLNA) ≤ 2%; NOEL (HRIPT)≤ 500 µg/cm2 • Category 1B – 2%; < EC3 (LLNA) < 50%; 500 µg/cm2 < NOEL (HRIPT) < 12 500 µg/cm2 • No alert found – EC3 (LLNA) ≥ 50%; Negative (LLNA); NOEL (HRIPT) > 12 500 µg/cm2

Each alert is associated with reaction mechanistic domain, e.g. Schiff base formation, Michael addition, Acylation, etc. and it is specified in the help file supporting the mechanism of interaction of the alerting group with skin proteins.
The profiler consists of structural and parametric (log KOW) requirements.

Relevance/Applicability to particular chemical classes

This profiler is applicable to those organic chemicals that have presence of at least one of the 135 protein binding alerts specified within the profiler. The presence of some protein binding alerts is conbined with specific logKow ranges, i.e. the profiler contains structural and parametric boundaries. The absence of a protein binding alert should not be taken as an absence of toxicity.

Approach used to develop the profiler - Concise but informative description of:

a) The aim of this profiler is to investigate the presence of alerts within the target molecules responsible for interaction with proteins and to provide potency classification of the target chemicals.

b) The profiler was developed from a mechanistic rationale that the molecular initiating event for skin sensitisation for low molecular weight chemicals is due to covalent binding of chemicals to proteins in the skin.

c) The training set of 517 chemicals was used to define the boundaries of the protein binding alerts for skin sensitisation according to GHS classification. Currently, there are 137 protein binding alerts classifying chemicals into GHS categories 1A and 1B. Distribution of the alerts across GHS categories is as follows: 87 alerting groups belong to Category 1A and 48 alerting groups belong to Category 1B.
This profiler accounts for incapability of some chemicals having an alert to interact with skin due to electronic and steric factors. This is explicitly defined by inhibition masks associated with some of the alerts.

d) Reference source

Summary description of profiles/alerts within the profiler

Profile/structural alert	Number of chemicals analysed	Number of chemicals associated with skin sensitisation
Skin sensitization Category 1A		
(Thio)Acyl and (thio)carbamoyl halides, cyanides, azides, etc.	6	2/6
1,2-Dicarbonyls, 1,3-Dicarbonyls	4	3/4
2-(Haloalkylidene)phenylhydrazines	0	0/0
Activated (di)aryl esters	2	2/2
Activated (thio)esters	0	0/0
Activated aryl and heteroaryl compounds	17	16/17
Activated Carbonyl compounds	2	2/2
Activated electrophilic ethenylarenes	0	0/0
Active cyclic agents	1	1/1
Alkene sultones	0	0/0
Allylic primary pseudohalides	0	0/0
Alpha- or beta-halo ethers	0	0/0
alpha, beta-Aldehydes	3	2/3

alpha, beta-carbonyl compounds with polarized double bond	23	19/23
alpha,beta-Carbonyl compounds with polarized triple bond	0	0/0
alpha-Activated benzyls	5	5/5
alpha-Activated haloalkanes	1	1/1
alpha-Ketoesters	0	0/0
Anhydrides (sulphur analogues of anhydrides)	4	4/4
Arene sulfinic acids	0	0/0
Aromatic carbonyl compounds	3	2/3
Aryliodonium salts	0	0/0
Azlactones and unsaturated lactone derivatives	4	4/4
Azocarbonamides	0	0/0
Azomethyne type compounds	0	0/0
Azomethynes with a sulfo leaving group	0	0/0
Azoxy compounds	0	0/0
Benzoyl phosphine oxides	0	0/0
beta-Lactams	1	1/1
Bifunctional alpha, beta-carbonyl containing compounds	1	1/1
Bis Aldehydes	5	4/5
Bis Epoxides	4	4/4
Carbodiimides	0	0/0
Conjugated systems with electron withdrawing groups	0	0/0
Diacyl peroxides, anhydrides (sulphur analogues of diacyl peroxides)	1	1/1
Diol epoxides	2	2/2
Dithiocarbamate salts	2	2/2
Dithiocarbamates	1	1/1
Dithioesters	0	0/0
Epoxides, Aziridines and Sulfuranes	2	2/2
Formaldehyde	4	4/4
Generated free radicals	10	9/10
Guanidines		
Halogenated five membered aromatic compounds	0	0/0
Halogenated isothiazolones	2	2/2
Halogenated nitroquinones	0	0/0
Heteroarene sulfenamides	0	0/0
Iodoalkynes	1	1/1
Isocyanates, Isothiocyanates	6	6/6

Isothiazolidin-3-ones (sulphur) and Benzoisothiazolinone	2	2/2
Isothiazolone derivatives	4	4/4
Ketenes	0	0/0
Lactones	2	2/2
Mercury compounds	0	0/0
Mustard compounds	0	0/0
N-Chloro-sulphonamides	1	1/1
N-Haloacylamides	0	0/0
Nitroalkenes	0	0/0
Nitrosoalkenes	4	3/4
N-nitroso compounds	9	9/9
N-nitroso_compounds	9	9/9
N-oxicarbonyl amides, N-Acyloxy-N-alkoxyamides	0	0/0
N-Sulfonylazomethynes	1	1/1
Organic sulfonyl azides	0	0/0
Organic thiosulfates	0	0/0
Phenyl carbonates	0	0/0
Phosphonyl halides or cyanide	0	0/0
Phosphoranylidene compounds	0	0/0
Polarised alkene - alkenyl pyridines, pyrazines, pyrimidines or triazines	2	2/2
Polarised Alkenes - sulfonates	0	0/0
Polarised Alkenes- sulfinyl	0	0/0
Polarised Alkenes- sulfones	1	1/1
Polarised alkynes, alkinyl pyridines, pyrazines, pyrimidines, triazines	0	0/0
Quinone(s)/imines, Quinone methide(s)/imines, Quinoide oxime structure, Nitroquinones, Naphtoquinone(s)/imines	91	56/91
Substituted benzyl benzoates	0	0/0
Sulfates	1	1/1
Sulfenyl halides	0	0/0
Sulfonates	4	3/4
Sulphonyl halides or cyanides	0	0/0
Sultones	0	0/0
Thiocyanates	2	2/2
Thio-lactones	0	0/0
Thiosulfinates	0	0/0
Thiosulfonates	0	0/0
Thiourea compounds	0	0/0
Vinyl-type compounds with electron-withdrawing groups	2	2/2

Skin sensitization Category 1B		
(Thio)Phosphates	1	1/1
1,2-Dicarbonyls, 1,3-Dicarbonyls_low activity	13	12/13
a,b-Unsaturated oximes	4	1/4
Activated (di)aryl esters_low activity	4	4/4
Activated (thio)esters_low ativity	1	1/1
Activated alkyl esters	7	7/7
Activated aryl and heteroaryl compounds_low activity	17	14/17
Active cyclic agents_low activity	2	2/2
Aldehydes	63	48/63
Alkyl diesters	0	0/0
Alkyl halides	23	22/23
alpha, beta-Aldehydes_low activity	14	12/14
alpha, beta-carbonyl compounds with polarized double bond_low reactivity	36	26/36
alpha,beta-Carbonyl compounds with polarized triple bond_low activity	1	1/1
alpha-activated acetates	8	7/8
alpha-Activated haloalkanes_low activity	2	2/2
alpha-Ketoesters_low activity	1	1/1
Amides	2	1/2
Azlactones and unsaturated lactone derivatives_low activity	4	4/4
Benzyl or phenethyl salicylates	1	0/1
Benzyl phenyl ethers	1	1/1
Bifunctional alpha, beta-carbonyl containing compounds_low activity	2	2/2
Carbamates	2	1/2
Carbenium ion	1	1/1
C-Nitroso compounds	2	0/2
Conjugated systems with electron withdrawing groups_low activity	3	2/3
Cyanoalkenes	1	1/1
Diacyl peroxides, anhydrides (sulphur analogues of diacyl peroxides)_low activity	0	0/0
Dithiocarbamate ester disulfides	2	2/2
Epoxides, Aziridines and Sulfuranes_low activity	20	16/20
Generated free radicals_low active	2	1/2
Hydroperoxides	37	36/37
Ketones	15	14/15

Lactones_low activity	1	1/1
N-Carbonyl heteroaryl amines	0	0/0
N-Carbonylsulfonamides	1	0/1
Nitrosoalkenes_low activity	1	1/1
N-nitroso compounds_low activity	0	0/0
N-nitroso_compounds_low activity	0	0/0
Phosphonates	0	0/0
Pyranones, Pyridones (and related nitrogen chemicals)	0	0/0
Pyrazolones and pyrazolidinones	2	1/2
Quinone(s)/imines, Quinone methide(s)/imines, Quinoide oxime structure, Nitroquinones, Naphtoquinone(s)/imines_low activity	9	5/9
Sulfates_low activity	1	1/1
Sulfonates_low activity	1	1/1
Sultones_low activity	0	0/0
Thiols, (poly)sulfides and dithiocarbamate ester disulfides	8	5/8
Total	571	459/571

Counter category: No alert found

Similar to other profilers

This profiler is similar to the general mechanistic ***Protein binding by OASIS*** and enpoitnt specific ***Protein binding alerts for skin sensitization by OASIS***. However, this profiler is endpoint specific and is designed to indicate chemicals could interact with proteins and could cause skin sensitisation. As might be expected there is significant overlap between the profilers (given that the MIE is the same); however, the structural alerts in the endpoint specific profiler are focussed on chemistry associated with covalent binding to skin proteins associated with skin allergy. In this respect the specificity of the profiler is codded by using a specific inhibition masks associated with some of structural alerts. As such, this profiler should be used not as a primary grouping method, but as a secondary method for refining the primary group of chemicals. As a result of this a stringent and more consistent group of chemical responsible for causing skin sensitisation effect could be obtained.

Short description of update version

SMARTS language for describing molecular patterns, i.e. structural boundaries, structural alerts has been implemented in OECD QSAR Toolbox 4.1. As a result "***Protein binding alerts for skin sensitization according to GHS***" has been implemented.
The updates in the profiler available in QSAR Toolbox 4.2 consist of:

- Addition of local training sets to the corresponding structural alerts including the following information:
  - Chemical ID (CAS, Name, SMILES)
  - Representative experimental data - in case of multiple data the worst case scenario or expert judgement is used
  - Metabolic activation

○ Bioassay (LLNA and HRIPT) ○ References
Disclaimer
The structural boundaries used to define the chemical classes (e.g. "Alcohol" – chemical class from "Organic functional group" profiler) or alerting groups responsible for the binding with biological macromolecules (e.g. "Aldehydes" – structural alert for protein binding), represent structural functionalities in the molecule which could be used for building chemical categories for subsequent data gap filling. They are not recommended to be used directly for prediction purposes (as SARs).

PROTEIN BINDING ALERTS FOR SKIN SENSITIZATION BY OASIS

All the chemicals of interest do not fire a known alert. The documentation does not provide an insight into the mispredictions but it provides only information on true positives and false positives. According to these statistics the Positive predictive value of the module = TP/(TP+FP) = 94% . According to the article by Urbisch *et al.* (2016) if this profiler is compared to LLNA assays then its Cooper's statistics are as follows: Sensitivity 67%, Specificity 82%, Accuracy 71%

References:

Urbisch *et al.* (2016) Peptide reactivity associated with skin sensitization: The QSAR Toolbox and TIMES compared to the DPRA. Toxicol *In vitro*. 2016 Aug;34:194-203. doi: 10.1016/j.tiv.2016.04.005.

About section of a profiler
Name of the profiler
Protein binding alerts for skin sensitization according to GHS
Developer; Donator; date; version
Developer: Laboratory of Mathematical Chemistry (LMC), Bourgas, Bulgaria ***Donator:*** L'Oreal, ExxonMobil, Procter & Gamble, Uniliver, Research Institute for Fragrance Materials (RIFM), Dow Chemical, Danish National Food Institute, Denmark ***Version:*** 1.6 December 2017
Relevance/Applicability to endpoint(s)
This profiler has been developed by industry consortia involving ExxonMobile, Procter&Gamble, Uniliver, Research Institute for Fragrance Materials (RIFM), Dow and Danish National Food Institute with the Laboratory of Mathematical Chemistry and the partnership of Dr D.Roberts, as a part of the TIMES model for predicting skin sensitisation. The profiler is intended to be used for the assessment of protein binding interaction of chemicals and especially interaction with skin proteins. The profiler has been developed based on mechanistic knowledge for skin sensitisation of dataset of 881 chemicals tested by Local Lymph Node Assay (LLNA) or Guinea Pig Maximization Test (GPMT). A list of 110 structural alerts has been derived, based on the mechanistic knowledge of training set

chemicals. The list of 110 structural alerts has been separated into 11 mechanistic domains. Each of the mechanistic domains has been separated into more than 2 mechanistic alerts. The profiling result outcome assigns a target to the corresponding structural alert, mechanistic alerts and domain.

Relevance/Applicability to particular chemical classes

This profiler is applicable to those organic chemicals that have presence of at least one of the 110 protein binding alerts specified within the profiler. The presence of protein binding alerts is not bounded with parametric ranges; it is based on structural boundaries only. The absence of a structural alert should not be taken as an absence of toxicity.

Approach used to develop the profiler - Concise but informative description of:

a) The aim of this profiler is to investigate the presence of alerts within the target molecules responsible for interaction with proteins and especially with skin proteins.

b) The profiler was developed from a mechanistic rationale that the molecular initiating event for skin sensitisation for low molecular weight chemicals is due to covalent binding of chemicals to proteins in the skin.

c) The profiler was developed from a dataset of 881 chemicals that have experimental skin sensitisation data based on LLNA or GPMT. A list of 110 structural alerts has been derived. This profiler accounts for incapability of some chemicals having an alert to interact with skin due to electronic and steric factors. This is explicitly defined by inhibition masks associated with some of the alerts.

d) Reference source

Summary description of profiles/alerts within the profiler

Profile/structural alert	Number of chemicals analysed	Number of chemicals associated with skin sensitisation
(Thio)Acetates	0	0/0
(Thio)Acyl and (thio)carbamoyl halides, cyanides	8	8/8
(Thio)Phosphates	3	3/3
1,2-Dicarbonyls and 1,3-Dicarbonyls	21	20/21
2-(Haloalkylidene)phenylhydrazines	0	0/0
Activated alkyl diesters	3	3/3
Activated alkyl esters and thioesters	4	3/4
Activated aryl and heteroaryl compounds	23	23/23
Activated (di)aryl esters	38	37/38
Activated (thio)esters	1	1/1
Activated Carbonyl compounds	3	3/3
Activated electrophilic ethenylarenes	0	0/0
Active cyclic agents	3	3/3
Aldehydes	75	69/75
Alkene sultones	0	0/0
Alkyl halides	25	23/25
Allyl and propargyl sulfate and sulfonate esters	1	1/1
Alpha- or beta-Halo ether	0	0/0

alpha, beta-Aldehydes	21	19/21
alpha, beta-Carbonyl compounds with polarized double bond	65	58/65
alpha,beta-Carbonyl compounds with polarized triple bond	3	3/3
alpha,beta-Unsaturated oximes	4	4/4
alpha-Activated acetates	8	7/8
alpha-Activated benzyls	6	6/6
alpha-Activated haloalkanes	8	7/8
alpha-Ketoesters	2	2/2
Amides	3	3/3
Anhydrides (sulphur analogues of anhydrides)	4	4/4
Arenesulfinic acids	0	0/0
Aromatic carbonyl compounds	10	10/10
Aryliodonium salts	0	0/0
Azlactones and unsaturated lactone derivatives	10	10/10
Azocarbonamides	0	0/0
Azomethyme type compounds	0	0/0
Azomethynes with a sulfo leaving group	1	1/1
Azoxy compounds	0	0/0
Benzoyl phosphine oxides	0	0/0
Benzyl or phenethyl salicylates	2	2/2
Benzyl phenyl ethers	1	1/1
beta-Lactams	1	1/1
Bifunctional alpha, beta-carbonyl containing compounds	5	4/5
Bis aldehydes	5	5/5
Carbamates	4	3/4
Carbenium ion	1	1/1
Carbodiimides	0	0/0
C-Nitroso compounds	2	0/2
Conjugated systems with electron withdrawing groups	6	6/6
Cyanoalkenes	3	3/3
Cyclopropenones	1	1/1
Diacyl peroxides, anhydrides (sulphur analogues of diacyl peroxides)	2	2/2
Di-substituted alpha,beta-unsaturated aldehydes	7	7/7
Dithiocarbamate salts	2	2/2
Dithiocarbamates	0	0/0
Dithioesters	4	4/4
Epoxides, Aziridines and Sulfuranes	33	30/33

Generated free radicals	19	19/19
Guanidines	1	1/1
Halogenated five membered aromatic compounds	0	0/0
Halogenated isothiazolones	2	2/2
Halogenated nitroquinones	0	0/0
Heteroarene sulfenamides	0	0/0
Hydroperoxides	58	52/58
Iodoalkynes	1	1/1
Isocyanates, Isothiocyanates	8	8/8
Isothiazolidin-3-ones (sulphur) and Isothiazolone derivatives	1	1/1
Isothiazolone derivatives	5	5/5
Ketenes	0	0/0
Ketones	18	15/18
Lactones	4	4/4
Mercury compounds	0	0/0
Mustard compounds	0	0/0
N-Carbonyl heteroaryl amines	0	0/0
N-Acylloxysuccinimides	0	0/0
N-Chloro-sulphonamides	1	1/1
N-Haloacylamides	1	1/1
Nitroalkenes	0	0/0
Nitrosoalkenes	5	5/5
N-Nitroso compounds (subsitution on N-atom)	3	3/3
N-Nitroso compounds (SN2, on sp3 carbon atoms)	3	3/3
N-Nitroso_compounds (central C atom)	3	3/3
N-oxicarbonyl amides, N-Acyloxy-N-alkoxyamides	0	0/0
N-Sulfonylazomethynes	1	1/1
Organic sulfonyl azides	0	0/0
Organic thiosulfates and thiosulfonates	1	1/1
Phenyl carbonates	6	6/6
Phoshoranylidene compounds	0	0/0
Phosphonates	0	0/0
Phosphonyl halides or cyanide	0	0/0
Polarised Alkene - alkenyl pyridines, pyrazines, pyrimidines or triazines	2	2/2
Polarised Alkenes- sulfinyl	1	1/1
Polarised Alkenes - sulfonates	0	0/0
Polarised Alkenes- sulfones	2	2/2
Polarised Alkynes, alkinyl pyridines, pyrazines, pyrimidines, triazines	0	0/0

Pyranones, Pyridones (and related nitrogen chemicals)	2	2/2
Pyrazolones and pyrazolidinones	2	2/2
Quinone methide(s)/imines, Quinoide oxime structure; Nitroquinones, Naphtoquinone(s)/imines	119	113/119
Substituted benzyl benzoates	1	1/1
Sulfates	2	2/2
Sulfenyl halides	0	0/0
Sulfonates	6	6/6
Sulphonyl halides or cyanides	1	1/1
Sultones	0	0/0
Thiocyanates	3	3/3
Thio-lactones	0	0/0
Thiols and disulfide compounds	16	15/16
Thiosulfinates	0	0/0
Thiosulfonates	0	0/0
Thiourea compounds	2	2/2
Vinyl type compounds with electron withdrawing groups	1	1/1
Total	734	690/734

Counter category: No alert found

Similar to other profilers

This profiler is similar to the general mechanistic ***Protein binding by OASIS***. However, this profiler is endpoint specific and is designed to indicate chemicals could interact with proteins and could cause skin sensitisation. As might be expected there is significant overlap between the profilers (given that the MIE is the same); however, the structural alerts in the endpoint specific profiler are focussed on chemistry associated with covalent binding to skin proteins associated with skin allergy. In this respect, the specificity of the profiler is codded by using a specific inhibition masks associated with some of structural alerts. As such, this profiler should be used not as a primary grouping method, but as a secondary method for refining the primary group of chemicals. As a result of this a stringent and more consistent group of chemical responsible for causing skin sensitisation effect could be obtained.

Short description of update version

SMARTS language for describing molecular patterns, i.e. structural boundaries, structural alerts has been implemented in OECD QSAR Toolbox 4.0. As a result, "***Protein binding alerts for skin sensitization by OASIS***" has been rewritten. Only small distinctions are expected in the profiling results between Toolbox v.3.4 and v 4.0 due to different 4 interpretation of the molecular structures, e.g. for heterocyclic/heteroaromatic compounds. Further general modifications are as follows:

1. Ketones - slight correction in the structural boundary - N-atom is removed
2. α-Haloalkenes (and related cyano, sulfate and sulphonate substituted chemicals) were renamed to Allyl and propargyl sulfate and sulfonate esters
3. The mechanistic alert named Electrostatic interaction of tetraalkylamonium ion with protein carboxylates was deleted

Modifications in OECD QSAR Toolbox 4.1 are as follows:

1. The structural boundaries of the category named "Activated aryl esters" were separated into 5 new categories:
   - Activated (di)aryl esters
   - Activated (thio)esters
   - Benzyl or phenethyl salicylates
   - Phenyl carbonates
   - Substituted benzyl benzoates
2. New categories were defined:
   - Bis aldehydes
   - Bifunctional alpha, beta - carbonyl containing compounds
   - Benzyl phenyl ethers
   - Iodoalkynes
   - Alpha - activated acetates
3. The category named "Vinyl type compounds with electron withdrawing groups" were split into 2 categories:
   - Vinyl type compounds with electron withdrawing groups
   - Azomethynes with a sulfo leaving group
4. The category Activated alkyl esters was renamed to Activated alkyl diesters
5. The category Organic thiosulfates was renamed to Organic thiosulfates and thiosulfonates due to addition of a new structural boundary
6. The category Pyranones, Pyridones (and related nitrogen chemicals) were added to the mechanistic alert named "Michael addition on quinoid type compounds"
7. Modifications were also done in the structural boundaries of the following categories based on expert judgement and available training set representatives:
   - Activated aryl and heteroaryl compounds
   - Lactones
   - Aldehydes
   - alpha - ketoesters
   - alpha,beta-Carbonyl compounds with polarised double bonds
   - Dithiocarbamates
   - alpha-Activated benzyls
   - alpha,beta-Aldehydes
   - Ketones
   - Cyanoalkenes
   - Aromatic carbonyl compounds
   - Activated alkyl esters and thioesters
   - 1,2-Dicarbonyls and 1,3-Dicarbonyls

The updates in the profiler available in QSAR Toolbox 4.2 consist of:

- Addition of local training sets to the corresponding structural alerts including the 5 following information:
  - Chemical ID (CAS, Name, SMILES)

○ Representative experimental data - in case of multiple data the worst case scenario or expert judgement is used ○ Metabolic activation ○ Bioassay (LLNA and GPMT) ○ References
Disclaimer
The structural boundaries used to define the chemical classes (e.g. "Alcohol" – chemical class from "Organic functional group" profiler) or alerting groups responsible for the binding with biological macromolecules (e.g. "Aldehydes" – structural alert for protein binding), represent structural functionalities in the molecule which could be used for building chemical categories for subsequent data gap filling. They are not recommended to be used directly for prediction purposes (as SARs).

PROTEIN BINDING BY OASIS

All the chemicals of interest do not fire a known alert. The documentation does not report statistics on predictivity. According to the article by Urbisch *et al.* (2016) if this profiler is compared to LLNA assays then its Cooper's statistics are as follows: Sensitivity 67%, Specificity 82%, Accuracy 71%.

Reference:

Urbisch *et al.* (2016) Peptide reactivity associated with skin sensitization: The QSAR Toolbox and TIMES compared to the DPRA. Toxicol *In vitro*. 2016 Aug;34:194-203. doi: 10.1016/j.tiv.2016.04.005.

About section of a profiler
Name of the profiler
Protein binding by OASIS
Developer; Donator; date; version
Developer: Laboratory of Mathematical Chemistry (LMC), Bourgas, Bulgaria ***Donator:*** Laboratory of Mathematical Chemistry (LMC), Bourgas, Bulgaria; L'Oréal; ExxonMobil; Procter & Gamble; Unilever; Research Institute for Fragrance Materials (RIFM), Dow Chemical, Danish National Food Institute, Denmark ***Version:*** 1.5 December 2017
Relevance/Applicability to endpoint(s)
The profiler is based on the rules defined in the OASIS TIMES models for Skin sensitisation (SS). It consists of 112 structural alerts related to interactions with proteins especially skin proteins and proteins such as topoisomerases, cellular protein adducts, etc. It is believed that positive results are result of interactions with proteins. The list of structural alerts has been separated into 11 mechanistic domains. Each of the mechanistic domains has been separated

into more than 2 mechanistic alerts. The profiling result outcome assigns a target to the corresponding structural alert, mechanistic alerts and domain.
Relevance/Applicability to particular chemical classes
This profiler is applicable to those organic chemicals that have presence of at least one of the 112 protein binding alerts specified within the profiler. The presence of protein binding alerts is not bounded with parametric ranges; it is based on structural boundaries only. The absence of a structural alert should not be taken as an absence of toxicity.
Approach used to develop the profiler - Concise but informative description of:
a) The overall rational: The aim of this profiler is to investigate the presence of alerts within the target molecules responsible for interaction with proteins.
b) The criteria or the method applied for analysing the training set/the pool of chemicals that inform the profiler: The profiler was developed from a mechanistic rationale that the molecular initiating event for skin sensitisation for low molecular weight chemicals is due to covalent binding of chemicals to proteins in the skin.
c) Source of the data/knowledge and total number of chemicals included in the analysis:
d) Literature references:
Summary description of profiles/alerts within the profiler
Summary list of the profiler categories is provided below**.**
Profile categories:
Isothiocyanates, Isocyanates
Carbodiimides
(Thio) Acetates
(Thio)Acyl and (thio)carbamoyl halides and cyanides
Anhydrides (sulphur analogues of anhydrides)
Azlactones and unsaturated lactone derivatives
Carbamates
Diacyl peroxides, anhydrides (sulphur analogues of diacyl peroxides)
N-Acylloxysuccinimides
N-Carbonyl heteroaryl amines
N-Carbonylsulfonamides
N-Haloacylamides
Phosphonyl halides or cyanides
Sulphonyl halides or cyanides
Thiosulfinates
Thiosulfonates
Amides
Dithiocarbamate salts
Dithiocarbamates
Dithioesters
Activated (di)aryl esters
Activated (thio)esters
Activated alkyl diesters
Benzyl or phenethyl salicylates
Phenyl carbonates
Substituted benzyl benzoates

Ketenes
Active cyclic agents
beta-Lactams
Cyclopropenones
Thio-lactones
Tetraalkylammonium ions
Guanidines
alpha,beta-Aldehydes
Lactones
Azoxy compounds
Activated electrophilic ethenylarenes
alpha,beta-Carbonyl compounds with polarized double bonds
alpha,beta-Carbonyl compounds with polarized triple bond
Bifunctional alpha, beta-carbonyl containing compounds
Conjugated systems with electron withdrawing groups
Cyanoalkenes
Nitroalkenes
Nitrosoalkenes
N-Sulfonylazomethynes
Phosphoranylidene compounds
alpha,beta-Unsaturated oximes
Polarised alkene - alkenyl pyridines, pyrazines, pyrimidines or triazines
Polarised Alkenes – sulfinyl
Polarised Alkenes – sulfonates
Polarised Alkenes – sulfones
Polarised alkynes – alkinyl pyridines, pyrazines, pyrimidines, triazines
Azocarbonamides
Pyranones, Pyridones (and related nitrogen chemicals)
Quinone methide(s)/imines; Quinoide oxime structure; Nitroquinones, Naphthoquinone(s)/imines
Alkene sultones
Azomethyme type compounds
Ketones
C-Nitroso compounds
Generated free radicals
Hydroperoxides
Organic peroxy compounds
Benzoyl phosphine oxides
1,2-Dicarbonyls and 1,3-Dicarbonyls
Di-substituted alpha,beta-unsaturated aldehydes
Activated Carbonyl compounds
Aldehydes
alpha-Ketoesters
Aromatic carbonyl compounds

Bis aldehydes
Pyrazolones and Pyrazolidinones
Carbenium ion
Mercury compounds
Allyl and propargyl sulfate and sulfonate esters
Thiols and disulfide compounds
Iodoalkynes
N-Nitroso compounds (SN2-Nucleophilic substitution at a Nitrogen atom)
N-oxicarbonyl amides, N-Acyloxy-N-alkoxyamides
(Thio)Phosphates
Alkyl halides
alpha-Activated haloalkanes
N-Nitroso compounds (SN2-Nucleophilic substitution at sp3 carbon atom)
Phosphonates
Sulfates
Sulfonates
N-Nitroso_compounds (Nucleophilic substitution at the central carbon atom of N-nitroso compounds)
Organic thiosulfates and thiosulfonates
alpha-Activated benzyls
Heteroarene sulfenamides
Organic sulfonyl azides
Epoxides, Aziridines and Sulfuranes
Isothiazolone derivatives
Mustards compounds
Sultones
N-Chloro-sulphonamides
2-(Haloalkylidene)phenylhydrazines
Activated alkyl esters and thioesters
alpha- or beta-Halo ethers
Benzyl phenyl ethers
Isothiazolidin-3-ones (sulphur) and Isothiazolone derivatives
Sulfenyl halides
Thiocyanates
Arenesulfinic acids
Thiourea compounds
Activated aryl and heteroaryl compounds
Halogenated five membered aromatic compounds
Halogenated nitroquinones
Aryliodonium salts
Halogenated isothiazolones
Azomethynes with a sulfo leaving group
Vinyl-type compounds with electron withdrawing groups
Total: 112 categories

Counter category: Not categorised

Similar to other profilers

This profiler is general mechanistic and it is similar to the ***Protein binding alerts for skin sensitization by OASIS*** and ***Protein binding by OECD***. As might be expected there is significant overlap between the profilers (given that the MIE is the same); however, the structural alerts in the general specific profiler are focussed on chemistry associated with covalent binding to skin proteins associated with skin allergy. In this respect, the specificity of the profiler is coded by using a specific inhibition masks associated with some of structural alerts. As such, this profiler should be used not as a primary grouping method, but as a secondary method for refining the primary group of chemicals. As a result of this a stringent and more consistent group of chemical responsible for causing skin sensitisation effect could be obtained.

Short description of update version

SMARTS language for describing molecular patterns, i.e. structural boundaries, structural alerts has been implemented in OECD QSAR Toolbox 4.0. As a result***, Protein binding by OASIS*** has been rewritten. Only small distinctions are expected in the profiling results between Toolbox v.3.4 and v 4.0 due to different interpretation of the molecular structures, e.g. for heterocyclic/heteroaromatic compounds.

Further general modifications are as follows:

1. Ketones - slight correction in the structural boundary - N-atom is removed
2. α-Haloalkenes (and related cyano, sulfate and sulphonate substituted chemicals) were renamed to Allyl and propargyl sulfate and sulfonate esters
3. The mechanistic alert named Electrostatic interaction of tetraalkylamonium ion with protein carboxylates was deleted
4. Removing of the rules defined in the OASIS TIMES models for Cromosomal aberration. Removed are as follows:

1	Carbamates
2	alpha,beta-Unsaturated Carbonyls and Related Compounds
3	Isothiocyanates
4	Isocyanates and Diisocyanates
5	Pyrazolone and Pyrazolidine-3,5-dione Derivatives
6	alpha,omega-Dihaloalkanes
7	Ethenyl Pyridines
8	Pyrimidines and Purines
9	Bipyridilium Herbicides
10	alpha,beta-Unsaturated Carboxylic Acids and Esters
11	Carboxylic Acid Anhydrides
12	Halogenated Vicinal Hydrocarbons
13	Heterocyclic Aromatic Amines
14	Substituted Phenols
15	Carboxylic Acid Amides
16	Arenesulfonamides
17	N-Substituted Aromatic Amines
18	Arenecarboxylic Acid Esters
19	Cyanohydrins

20	Substituted Anilines
21	Gallic Acid Esters
22	Benzoquinoline and Acridine derivatives
23	N-Alkyl-N-nitrosocarbamates
24	N-Nitrosoamine Derivatives
25	Alkylated nitrosoureas and nitrosoguanidines
26	Dialkyl Alkylphosphonates
27	Hexahydrotriazine Derivatives
28	Sterically Hindered Piperidine Derivatives
29	Hydroxylated Phenols
30	Propargyl Alcohol Derivatives
31	Quinoneimine

Modifications in OECD QSAR Toolbox 4.1 are as follows:

1. The structural boundaries of the category named "Activated aryl esters" were separated into 5 new categories:
   - Activated (di)aryl esters
   - Activated (thio)esters
   - Benzyl or phenethyl salicylates
   - Phenyl carbonates
   - Substituted benzyl benzoates
2. New categories were defined:
   - Bis aldehydes
   - Bifunctional alpha, beta - carbonyl containing compounds
   - Benzyl phenyl ethers
   - Iodoalkynes
   - Alpha - activated acetates
3. The category named "Vinyl type compounds with electron withdrawing groups" were split into 2 categories:
   - Vinyl type compounds with electron withdrawing groups
   - Azomethynes with a sulfo leaving group
4. The category Activated alkyl esters was renamed to Activated alkyl diesters
5. The category Organic thiosulfates was renamed to Organic thiosulfates and thiosulfonates due to addition of a new structural boundary
6. The category Pyranones, Pyridones (and related nitrogen chemicals) were added to the mechanistic alert named "Michael addition on quinoid type compounds"
7. Modifications were also done in the structural boundaries of the following categories based on expert judgement and available training set representatives:
   - Activated aryl and heteroaryl compounds
   - Lactones
   - Aldehydes
   - alpha - ketoesters
   - alpha,beta-Carbonyl compounds with polarised double bonds
   - Dithiocarbamates
   - alpha-Activated benzyls

- alpha,beta-Aldehydes - Conjugated systems with electron withdrawing groups - Cyanoalkenes - Aromatic carbonyl compounds - Activated alkyl esters and thioesters - 1,2-Dicarbonyls and 1,3-Dicarbonyls 8. Restore of the mechanistic alert named Electrostatic interaction of tetraalkylamonium ion with protein carboxylates
Disclaimer
The structural boundaries used to define the chemical classes (e.g. "Alcohol" – chemical class from "Organic functional group" profiler) or alerting groups responsible for the binding with biological macromolecules (e.g. "Aldehydes" – structural alert for protein binding), represent structural functionalities in the molecule which could be used for building chemical categories for subsequent data gap filling. They are not recommended to be used directly for prediction purposes (as SARs).

PROTEIN BINDING BY OECD

All the chemicals of interest do not fire a known alert. The documentation does not report statistics on predictivity. According to the article by Urbisch *et al.* (2016) if this profiler is compared to LLNA assays then its Cooper's statistics are as follows: Sensitivity 63%, Specificity 85%, Accuracy 69%

Reference:

Urbisch *et al.* (2016) Peptide reactivity associated with skin sensitization: The QSAR Toolbox and TIMES compared to the DPRA. Toxicol *In vitro*. 2016 Aug;34:194-203. doi: 10.1016/j.tiv.2016.04.005.

About section of a profiler
Name of the profiler
Protein binding by OECD
Developer; Donator; date; version
Developer: School of Pharmacy and Chemistry, Liverpool John Moores University, UK ***Donator:*** European Chemicals Agency (ECHA); Organisation for Economic Co-operation and Development (OECD) ***Version:*** 2.3 December 2016
Relevance/Applicability to endpoint(s)
This profiler is intended to be used for the assessment of endpoints in which covalent binding to a protein has been shown to be the molecular initiating event for low molecular weight chemicals. The profiler has been developed from mechanistic knowledge of the electrophilic chemistry of covalent protein binding for direct acting electrophiles only – importantly it has

been developed from a systematic review of the literature and not from the analysis of a single toxicological dataset.
Relevance/Applicability to particular chemical classes
This profiler is applicable only to organic chemicals that have a molecular weight less than 1000 g/mol. It is applicable only to the chemical classes for which it contains structural alerts; the absence of a structural alert should not be taken as an absence of toxicity. This profiler contains structural alerts for direct acting electrophiles only – oxidation and/or metabolism are not accounted for (appropriate Toolbox simulators should be applied, if required).
Approach used to develop the profiler - Concise but informative description of:
a) The aim of the profiler was to identify structural alerts associated with organic, low molecular weight chemicals capable of forming covalent bonds with a protein. The structural alerts were derived from knowledge of the molecular initiating event - covalently binding to a protein. It was developed from a systematic review of the literature, rather than from the analysis of a single toxicological dataset.
b) The profiler was developed from a mechanistic rationale that the molecular initiating event for covalent bond formation with proteins. Importantly, this was achieved by reviewing the literature relating to the chemistry, rather than an analysis of toxicological datasets.
c) The profiler was developed from an extensive review of the literature relating to the chemistry of covalent bond formation with a protein. A full list of the literature included can be found in the reference listed in section d.
d) An overview of the mechanistic chemistry and underlying principles of the structural alerts within this profiler can be found in: Enoch *et al* (2010) *A review of the electrophilic reaction chemistry involved in covalent protein binding*. Critical Reviews in Toxicology, 41, p783-802
Summary description of profiles/alerts within the profiler
It is not possible to provide metrics relating to this profiler as it was not developed from an analysis of toxicological datasets. It was developed from an extensive review of the chemistry related to the formation of a covalent bond between a low molecular weight chemical and a protein.
Similar to other profilers
A number of related endpoint specific profilers exist in the OECD QSAR Toolbox relating to genotoxicity. The protein binding by OECD profiler should be used first, with endpoint specific profilers (which have been developed from an analysis of toxicological data) being used to sub-categorise, where possible. This profiler contains structural alerts for direct acting electrophiles only – oxidation and/or metabolism are not accounted for (appropriate Toolbox simulators should be applied, if required).
Short description of update version
SMARTS language for describing molecular patterns, i.e. structural boundaries, structural alerts has been implemented in OECD QSAR Toolbox 4.0. As a result, ***Protein binding by OECD*** has been rewritten but without modifying the knowledge and/or the logic, it is based on. Only small distinctions are expected in the profiling results between Toolbox v.3.4 and v4.0 due to different interpretation of the molecular structures, e.g. for heterocyclic/heteroaromatic compounds.

Further general modifications are associated with the new 2D editor which allows the structural boundaries to be coded more accurately according to the descriptions of the categories. Examples for categories with possible discrepancies between TB 3.4 and TB 4.0: Acetates; Allyl acetates and related chemicals.
Disclaimer
The structural boundaries used to define the chemical classes (e.g. "Alcohol" – chemical class from "Organic functional group" profiler) or alerting groups responsible for the binding with biological macromolecules (e.g. "Aldehydes" – structural alert for protein binding), represent structural functionalities in the molecule which could be used for building chemical categories for subsequent data gap filling. They are not recommended to be used directly for prediction purposes (as SARs).

PROTEIN BINDING POTENCY Cys (DPRA 13%)

All the chemicals of interest are characterised by a DPRA less than 9%. The documentation does not report statistics on predictivity but an article by Dimitrov *et al.* (2016) reports the following performance for a threshold = 13.3%: Concordance = 84%, Sens itivity = 96%, Specificity = 43%.

Reference:

Dimitrov *et al.* (2016) Accounting for data variability, a key factor in *in vivo*/*in vitro* relationships: application to the skin sensitization potency (*in vivo* LLNA versus *in vitro* DPRA) example. J Appl Toxicol. 2016 Dec;36(12):1568-1578.

About section of a profiler
Name of the profiler
Protein binding potency Cys (DPRA 13%)
Developer; Donator; date; version
Developer: Laboratory of Mathematical Chemistry (LMC), Bourgas, Bulgaria ***Donator:*** Natsch *et al.*, Urbisch *et al.*, Jaworska *et al.* ***Version:*** 1.0 December 2016
Relevance/Applicability to endpoint(s)
This profile is built in relation with the implementation of the adverse outcome pathway (AOP) for skin sensitisation. It is developed on the base of data derived from Direct Peptide Reactivity Assay (DPRA). The DPRA is a reactivity assay which evaluates the ability of chemicals to react with proteins. As model peptides are used reduced glutathione and two synthetic peptides – lysine and cysteine. The reaction time for both lysine and cysteine is 24 hours. The peptide reactivity is reported as percent peptide depletion. The profile contains 77 structural alerts extracted from about 229 chemicals with experimentally measured cysteine depletion values. The set of 77 structural alerts are separated into three potency categories: DPRA above 21% (DPRA 13%), DPRA less than 9% (DPRA 13%) and Grey

zone 9-21% (DPRA 13%). Classification of potency categories is based on analysis published in a collaboration with L`Oreal (Dimitrov *et al.*, 2016).

Relevance/Applicability to particular chemical classes

This profiler is applicable to those organic chemicals that have presence of a functional group reacting with the cysteine residue. The presence of an alert is not bounded with parametric ranges; it is based on structural boundaries only.

Approach used to develop the profiler - Concise but informative description of:

a) The aim of the profiler is to investigate the chemicals for presence of functional group able to interact with cysteine peptide and to provide information about the potency of the interaction.

b) The profiler was developed on a basis of experimental data for reactivity toward model peptide (cysteine) measured as percent peptide depletion. The data has been obtained by measuring the covalent binding of the target chemicals with the thio group of cysteine (Cys).

c) The profiler was based on a dataset of 229 chemicals with experimental Cys depletion values. A list of 77 structural alerts has been derived. The structural alerts have been separated into three potency categories based on specific peptide depletion ranges. Each alert was associated by a local list of training set chemicals.

d) Literature references:

- Gerberick, G.F., Vassallo, J.D., Bailey, R.E., Chaney, J.G., Morrall, S.W. and Lepoittevin, J.P. 2004. Development of a peptide reactivity assay for screening contact allergens. Toxicol. Sci. 81: 332-343.
- Natsch, A. and Gfeller, H. 2008. LC-MS-based characterization of the peptide reactivity of chemicals to improve the *in vitro* prediction of the skin sensitisation potential. Toxicol. Sci. 106: 464-478.
- Natsch, A., Emter, R., Gfeller, H., Haupt, T. and Ellis, G. 2015. Predicting skin sensitizer potency based on *in vitro* data from KeratinoSens and kinetic peptide binding: global versus domain-based assessment. Toxicol. Sci. 143(2), 319-332.
- Jaworska, J., Natsch, A., Ryan, C., Strickland, J., Ashikaga, T., Miyazawa, M. 2015. Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment: a decision support system for quantitative weight of evidence and adaptive testing strategy. Arch Toxicol, 2355-2383.
- Urbisch, D., Mehling, A., Guth, K., Ramirez, T., Honarvar, N., Kolle, S., Landsiedel, R., Jaworska, J., Kern, P., Gerberick, F., Natsch, A., Emter, R., Ashikaga, T., Miyazawa, M., Sakaguchi, H. 2015. Assessing skin sensitization hazard in mice and men using non-animal test methods. Regulatory Toxicology and Pharmacology 71: 337-351.
- S. Dimitrov, A. Detroyer, C. Piroird, C. Gomes, J. Eilstein, T. Pauloin, C. Kuseva, H. Ivanova, I. Popova, Y. Karakolev, S. Ringeisses, O. Mekenyan, Accounting for data variability, a key factor in *in vivo*/*in vitro* relationships: application to the skin sensitization potency (*in vivo* LLNA versus *in vitro* DPRA) example. J Appl Toxicol, 2016, DOI 10.1002/jat.3318

Summary description of profiles/alerts within the profiler

Profile/structural alert	Phys-chem parameter
DPRA above 21% (DPRA 13%)	
1,2- and 1,3-Diketones (reactive)	No parameter

1,2-Dihaloalkanes with Other Electron-Withdrawing Substituents	No parameter
2,4-Dinitrohaloarenes and 2,4-Dinitrophenyl thiocyanates	No parameter
Abietic acid	No parameter
Activated 1,3,5-triazine derivatives	No parameter
Alkyl alkanesulfonates	No parameter
alpha,beta-Unsaturated compounds with polarized triple bonds	No parameter
Aminophenol derivatives (reactive)	No parameter
Benzisothiazolinone derivatives	No parameter
Benzyl halides	No parameter
Branched acyl halides	No parameter
Conjugated alpha, beta-unsaturated aldehydes	No parameter
Conjugated alpha, beta-unsaturated esters (reactive)	No parameter
Conjugated alpha, beta-unsaturated ketones (reactive)	No parameter
Cyclopropenones	No parameter
Diacylperoxides	No parameter
Dialkylsulfates	No parameter
Disulfides	No parameter
Epoxides	No parameter
Ethylenediamine and polyethylene amines (reactive)	No parameter
Glycidyl ethers (epoxyethers)	No parameter
Halogenated isothiazolones	No parameter
Heterocyclic substituted urea compound with formaldehyde-releasing activity	No parameter
Hydroquinone and catechol derivatives (reactive)	No parameter
Isocyanates and Isothiocyanates	No parameter
Isothiazolinone derivatives	No parameter
m- and o-Phenylenediamine derivatives	No parameter
Maleic anhydride derivatives	No parameter
Non-Conjugated carboxylic acids and esters (reactive)	No parameter
Non-Conjugated monoaldehydes (reactive)	No parameter
N-Substituted aromatic amides (reactive)	No parameter
Oxazolones	No parameter
p-Phenylenediamine derivatives	No parameter
Primary iodoalkanes	No parameter
Saturated dialdehydes	No parameter
Squaric acid	No parameter
Substituted nitrosoarenes	No parameter
Thiols (reactive)	No parameter
Vinyl pyridines	No parameter
Vinylene 1,2-biscarboxylates	No parameter
DPRA less than 9% (DPRA 13%)	
1,1-Dihaloethenes	No parameter
1,2- and 1,3-Diketones (non reactive)	No parameter
5-pyrazolone derivatives	No parameter

Alcohols	No parameter
Alkanes	No parameter
alpha alkyl cinnamaldehydes	No parameter
Aminophenol derivatives (non reactive)	No parameter
Anionic surfactants	No parameter
Cationic surfactants	No parameter
Conjugated alpha, beta-unsaturated esters (non reactive)	No parameter
Conjugated alpha, beta-unsaturated ketones (non reactive)	No parameter
Coumarin derivatives	No parameter
Cyclic acid anhydrides (non reactive)	No parameter
Ethylenediamine and polyethylene amines (non reactive)	No parameter
Hydroquinone and catechol derivatives (non reactive)	No parameter
Mercaptoalcohols	No parameter
Mono-halo arenes	No parameter
No protein binding alert	No parameter
Non-Conjugated carboxylic acids and esters (non reactive)	No parameter
Non-Conjugated monoaldehydes (non reactive)	No parameter
N-Substituted aromatic amides (non reactive)	No parameter
Other alpha, beta-unsaturated compounds with polarized double bonds (non reactive)	No parameter
p-Aminoarene Sulfonamides	No parameter
Sulfanilic acid derivatives	No parameter
Thiols (non reactive)	No parameter
Vaniline derivatives	No parameter
Grey zone 9-21% (DPRA 13%)	
1,3-Diketones	No parameter
alpha, beta-unsaturated acids	No parameter
Conjugated alpha, beta-unsaturated ketones (Grey zone)	No parameter
Cyclic acid anhydrides (Grey zone)	No parameter
Hydroquinone and catechol derivatives (Grey zone)	No parameter
Lactones fused to aromatic rings	No parameter
N,N-Dialkyl-alpha,omega-Alkanediamines	No parameter
Non-Conjugated carboxylic acids and esters (Grey zone)	No parameter
Non-Conjugated monoaldehydes (Grey zone)	No parameter
Polysorbates	No parameter
Primary haloalkanes with short alkyl chain	No parameter
Total: 77 categories	
Counter category: Out of mechanistic domain	
Similar to other profilers	
The profiler is similar to the ***Protein binding potency Lys (DPRA 13%)*** and ***Protein binding potency*** profilers. All three profilers are focused on possibility of chemicals to interact with proteins on the *in chemico* reactivity level and may provide indication for protein binding potency of chemicals. In this respect, ***Protein binding potency Cys (DPRA 13%)*** profiler should be used not as a primary grouping method, but as a secondary method for refining the	

primary group of chemicals. As a result of this a stringent and more consistent group of chemical responsible for interaction with cell proteins could be obtained.
Short description of update version
SMARTS language for describing molecular patterns, i.e. structural boundaries, structural alerts has been implemented in OECD QSAR Toolbox 4.0. As a result ***Protein binding potency Cys (DPRA 13%)*** has been rewritten. Distinctions are expected in the profiling results between Toolbox v.3.4 and v.4.0 due to: • Different thresholds used for classification of chemicals – in Toolbox v.3.4 classification of potency categories is as follows: Low reactive (cysteine depletion = 5-40%), Moderate reactive (cysteine depletion = 40-80%), High reactive (cysteine depletion > 80%) while in Toolbox v.4.0 classification of potency categories is as follows: DPRA above 21% (DPRA 13%), DPRA less than 9% (DPRA 13%) and Grey zone 9-21% (DPRA 13%); • Different number of structural alerts – the profile contains 32 structural alerts in Toolbox v.3.4 and 77 structural alerts in Toolbox v.4.0; • Different number of chemicals used to extract the structural alerts – 112 chemicals for structural alerts implemented in Toolbox v.3.4 and 229 chemicals used for Toolbox v.4.0; • Different interpretation of the molecular structures, e.g. for heterocyclic/heteroaromatic compounds.
Disclaimer
The structural boundaries used to define the chemical classes (e.g. "Alcohol" – chemical class from "Organic functional group" profiler) or alerting groups responsible for the binding with biological macromolecules (e.g. "Aldehydes" – structural alert for protein binding), represent structural functionalities in the molecule which could be used for building chemical categories for subsequent data gap filling. They are not recommended to be used directly for prediction purposes (as SARs).

PROTEIN BINDING POTENCY Lys (DPRA 13%)

All the chemicals of interest are characterised by a DPRA less than 9%. The documentation does not report statistics on predictivity but an article by Dimitrov *et al.* (2016) reports the following performance for a threshold = 13.3%: Concordance = 84%, Sensitivity = 96%, Specificity = 43%.

Reference:

Dimitrov *et al.* (2016) Accounting for data variability, a key factor in *in vivo*/*in vitro* relationships: application to the skin sensitization potency (*in vivo* LLNA versus *in vitro* DPRA) example. J Appl Toxicol. 2016 Dec; 36(12):1568-1578.

About section of a profiler
Name of the profiler
Protein binding potency Lys (DPRA 13%)
Developer; Donator; date; version
Developer: Laboratory of Mathematical Chemistry (LMC), Bourgas, Bulgaria

Donator:
Natsch *et al.*, Urbisch *et al.*, Jaworska *et al.*

Version: 1.0
December 2016

Relevance/Applicability to endpoint(s)

This profile is built in relation with the implementation of the adverse outcome pathway (AOP) for skin sensitisation. It is developed on the base of data derived from Direct Peptide Reactivity Assay (DPRA). The DPRA is a reactivity assay which evaluates the ability of chemicals to react with proteins. As model peptides are used reduced glutathione and two synthetic peptides – lysine and cysteine. The reaction time for both lysine and cysteine is 24 hours. The peptide reactivity is reported as percent peptide depletion. The profile contains 73 structural alerts extracted from about 228 chemicals with experimentally measured lysine depletion values. The set of 73 structural alerts are separated into three potency categories: DPRA above 21% (DPRA 13%), DPRA less than 9% (DPRA 13%) and Grey zone 9-21% (DPRA 13%). Classification of potency categories is based on analysis published in collaboration with L`Oreal (Dimitrov *et al.*, 2016).

Relevance/Applicability to particular chemical classes

This profiler is applicable to those organic chemicals that have presence of a functional group reacting with the lysine residue. The presence of an alert is not bounded with parametric ranges; it is based on structural boundaries only.

Approach used to develop the profiler - Concise but informative description of:

a) The aim of the profiler is to investigate the chemicals for presence of functional group able to interact with Lysine peptide and to provide information about the potency of the interaction.

b) The profiler was developed on a basis of experimental data for reactivity toward model peptide measured as percent peptide depletion. The data has been obtained by measuring the covalent binding of the target chemicals with the amino group of lysine (Lys).

c) The profiler was based on a dataset of 228 chemicals with experimental Lys depletion values. A list of 73 structural alerts has been derived. The structural alerts have been separated into three potency categories based on specific peptide depletion ranges. Each alert was associated by a local list of training set chemicals.

d) Literature references:

- Gerberick, G.F., Vassallo, J.D., Bailey, R.E., Chaney, J.G., Morrall, S.W. and Lepoittevin, J.P. 2004. Development of a peptide reactivity assay for screening contact allergens. Toxicol. Sci. 81: 332-343.
- Natsch, A. and Gfeller, H. 2008. LC-MS-based characterization of the peptide reactivity of chemicals to improve the *in vitro* prediction of the skin sensitisation potential. Toxicol. Sci. 106: 464-478.
- Natsch, A., Emter, R., Gfeller, H., Haupt, T. and Ellis, G. 2015. Predicting skin sensitizer potency based on *in vitro* data from KeratinoSens and kinetic peptide binding: global versus domain-based assessment. Toxicol. Sci. 143(2), 319-332.
- Jaworska, J., Natsch, A., Ryan, C., Strickland, J., Ashikaga, T., Miyazawa, M. 2015. Bayesian integrated testing strategy (ITS) for skin sensitization potency assessment:

a decision support system for quantitative weight of evidence and adaptive testing strategy. Arch Toxicol, 2355-2383.

- Urbisch, D., Mehling, A., Guth, K., Ramirez, T., Honarvar, N., Kolle, S., Landsiedel, R., Jaworska, J., Kern, P., Gerberick, F., Natsch, A., Emter, R., Ashikaga, T., Miyazawa, M., Sakaguchi, H. 2015. Assessing skin sensitization hazard in mice and men using non-animal test methods. Regulatory Toxicology and Pharmacology 71: 337-351.
- S. Dimitrov, A. Detroyer, C. Piroird, C. Gomes, J. Eilstein, T. Pauloin, C. Kuseva, H. Ivanova, I. Popova, Y. Karakolev, S. Ringeisses, O. Mekenyan, Accounting for data variability, a key factor in *in vivo*/*in vitro* relationships: application to the skin sensitization potency (*in vivo* LLNA versus *in vitro* DPRA) example. J Appl Toxicol, 2016, DOI 10.1002/jat.3318

Summary description of profiles/alerts within the profiler

Profile/structural alert	Phys-chem parameter
DPRA above 21% (DPRA 13%)	
Activated 1,3,5-triazine derivatives	No parameter
Allyl glycidyl and benzyl glycidyl ethers (reactive)	No parameter
Aminophenol derivatives (reactive)	No parameter
Benzyl halides	No parameter
Conjugated alpha,beta-unsaturated aldehydes (reactive)	No parameter
Conjugated alpha, beta-unsaturated esters (reactive)	No parameter
Cyclic acid anhydrides	No parameter
Diacylperoxides	No parameter
Halogenated isothiazolone derivatives	No parameter
Hydroxybenzene derivatives and quinones (reactive)	No parameter
Isocyanates and Isothiocyanates	No parameter
Lactones fused to aromatic rings	No parameter
Maleic anhydride derivatives	No parameter
Nitroaniline derivatives	No parameter
Non-alpha,beta-conjugated monoaldehydes (reactive)	No parameter
Non-Conjugated carboxylic acids and esters (reactive)	No parameter
Non-conjugated mono- and diketones (reactive)	No parameter
Oxazolone derivatives	No parameter
Phenylenediamine derivatives (reactive)	No parameter
Saturated dialdehydes	No parameter
Vinylene 1,2-biscarboxylates	No parameter
DPRA less than 9% (DPRA 13%)	
1,1-Dihaloethenes	No parameter
1,2-Dihaloalkanes with Other Electron-Withdrawing Substituents	No parameter
5-pyrazolone derivatives	No parameter
Alcohols	No parameter
Alkanes	No parameter
Allyl glycidyl and benzyl glycidyl ethers (non reactive)	No parameter
Aminophenol derivatives (non reactive)	No parameter

Amphoteric surfactants	No parameter
Anionic surfactants	No parameter
Cationic surfactants	No parameter
Conjugated alpha,beta-unsaturated aldehydes (non reactive)	No parameter
Conjugated alpha, beta-unsaturated esters (non reactive)	No parameter
Conjugated alpha, beta-unsaturated ketones (non reactive)	No parameter
Coumarin derivatives	No parameter
Cyclopropenones	No parameter
Ethylenediamine, Polyethylene Amines and N,N-Dialkyl-alpha,omega Alkanediamines	No parameter
Guanidines	No parameter
Heterocyclic substituted urea compound with formaldehyde-releasing activity	No parameter
Hydroxybenzene derivatives and quinones (non reactive)	No parameter
Isothiazolone derivatives	No parameter
Mono-halo arenes	No parameter
N-Acylsulfonamides	No parameter
No protein binding alert	No parameter
Non-alpha,beta-conjugated monoaldehydes (non reactive)	No parameter
Non-Conjugated carboxylic acids and esters (non reactive)	No parameter
Non-conjugated mono- and diketones (non reactive)	No parameter
Nonionic surfactants	No parameter
N-substituted aromatic amides	No parameter
p-Aminoarene sulfonamides	No parameter
Phenylenediamine derivatives (non reactive)	No parameter
Squaric acid derivatives	No parameter
Straight-chain primary haloalkanes	No parameter
Substituted 1,4-phenylenediamines and 4-aminophenyl ethers	No parameter
Substituted nitrosoarenes	No parameter
Sulfanilic Acid Derivatives	No parameter
Thiols and disulfides (non reactive)	No parameter
Vaniline derivatives	No parameter
Vinyl pyridines	No parameter
Grey zone 9-21% (DPRA 13%)	
Alkyl alkanesulfonates and dilakylsulfates	No parameter
Allyl glycidyl and benzyl glycidyl ethers (Grey zone)	No parameter
Aminophenol derivatives (Grey zone)	No parameter
Branched acyl halides	No parameter
Conjugated alpha,beta-unsaturated aldehydes (Grey zone)	No parameter
Conjugated alpha,beta-unsaturated esters (Grey zone)	No parameter
Conjugated alpha, beta-unsaturated ketones (Grey zone)	No parameter
Halonitrobenzenes	No parameter
Non-alpha,beta-conjugated monoaldehydes (Grey zone)	No parameter
Non-Conjugated carboxylic acids and esters (Grey zone)	No parameter

Non-conjugated mono- and diketones (Grey zone)	No parameter
Other quinoid structures (Grey zone)	No parameter
Phenylenediamine derivatives (Grey zone)	No parameter
Thiols and disulfides (Grey zone)	No parameter
Total: 73 categories	
Counter category: Out of mechanistic domain	
Similar to other profilers	
The profiler is similar to the ***Protein binding potency Cys (DPRA 13%)*** and ***Protein binding potency*** profilers. All three profilers are focused on possibility of chemicals to interact with proteins on the *in chemico* reactivity level and may provide indication for protein binding potency of chemicals. In this respect, ***Protein binding potency Lys (DPRA 13%)*** profiler should be used not as a primary grouping method, but as a secondary method for refining the primary group of chemicals. As a result of this a stringent and more consistent group of chemical responsible for interaction with cell proteins could be obtained.	
Short description of update version	
SMARTS language for describing molecular patterns, i.e. structural boundaries, structural alerts has been implemented in OECD QSAR Toolbox 4.0. As a result, ***Protein binding potency Lys (DPRA 13%)*** has been rewritten. Distinctions are expected in the profiling results between Toolbox v.3.4 and v.4.0 due to: • Different thresholds used for classification of chemicals – in Toolbox v.3.4 classification of potency categories is as follows: Low reactive (lysine depletion = 5-40%), Moderate reactive (lysine depletion = 40-80%), High reactive (lysine depletion > 80%) while in Toolbox v.4.0 classification of potency categories is as follows: DPRA above 21% (DPRA 13%), DPRA less than 9% (DPRA 13%) and Grey zone 9-21% (DPRA 13%); • Different number of structural alerts – the profile contains 24 structural alerts in Toolbox v.3.4 and 73 structural alerts in Toolbox v.4.0; • Different number of chemicals used to extract the structural alerts – 110 chemicals for structural alerts implemented in Toolbox v.3.4 and 228 chemicals used for Toolbox v.4.0; • Different interpretation of the molecular structures, e.g. for heterocyclic/heteroaromatic compounds.	
Disclaimer	
The structural boundaries used to define the chemical classes (e.g. "Alcohol" – chemical class from "Organic functional group" profiler) or alerting groups responsible for the binding with biological macromolecules (e.g. "Aldehydes" – structural alert for protein binding), represent structural functionalities in the molecule which could be used for building chemical categories for subsequent data gap filling. They are not recommended to be used directly for prediction purposes (as SARs).	

3. Intracellular concentrations in *in vitro* assays - Introduction to Biokinetic Modelling

Effective concentrations determined in *in vitro* toxicological assays are routinely based on a range of nominal treatment concentrations. However, the use of the nominal treatment concentration as the driving concentration for observed toxicity *in vitro* does not account for factors that reduce the free concentration within the assay and determine the true effective concentration. For cell based assay systems, this will dictate the concentration available for distribution into the cell, and subcellular compartments, and so determine the driving concentration at the target site mediating toxicity. In order to more accurately translate concentration driven toxicity from *in vitro* to *in vivo*, it is necessary to correct the nominal effect concentration, accounting for the distribution of the compound within the assay system. Factors to be considered in modelling this *in vitro* distribution include binding to the plastics used in assays, exchange at the interface between culture media and the air in the culture vessel, and binding to components that may be included in the culture media (e.g. lipids and proteins originating from fetal bovine serum, FBS); the modelling of these processes is termed biokinetics (Blaauboer 2010).

A number of biokinetic models have been published that account for some or all of the factors listed above in cell based assays. Armitage and colleagues (Armitage, Wania *et al.* 2014) published a model framework to predict intracellular concentrations, correcting for some of the distribution factors in monolayer cell culture. This steady-state framework assumes instantaneous partitioning between media, headspace, serum-lipids, serum-proteins, dissolved organic material, and the cultured cell volume. However, a critical assumption of the Armitage model is that the test compounds are neutral or not significantly ionised under the conditions of the *in vitro* assay. This assumption of neutrality was, to some degree, addressed in the model developed by Fischer *et al* (Fischer, Henneberger *et al.* 2017) where the authors adopted the same steady-state assumption, but excluded the partitioning of compound into the headspace. The proposed model incorporated separate partition constants for both the ionised and unionised fraction of test compound, determining the fraction ionised assuming a uniform pH=7.4 throughout the test system. This neglects the differential ionisation potential between the culture media and intracellular water resulting from their differing pH. Furthermore, the interior of the cell itself is not a uniform environment, the microenvironment of specific organelles being maintained at pH specific to their function (i.e. lysosomes (pH≈4.5), mitochondria (pH≈8), cytosol (pH≈7). Indeed, the differential ionisation of compounds between organelles and intracellular water can result in the preferential sequestration of compounds within organelles; a phenomena commonly known by the misnomer ‘ion-trapping’ (Kazmi, Hensley *et al.* 2013). It is also critical to note that differences in the intrinsic permeability of the unionised/ionised form are not the only factors determining the distribution of ionised compound into cells. The potent ial difference maintained across the cell membrane, membrane potential (mV) can actively promote the uptake or exclusion of ionised compounds from the cell interior and can vary significantly between cell types.

Steady-state Biokinetic Model 2D Monolayer Cell Culture

Given that compounds in this read-across are monoprotic acids, significanlty ionised at physiological pH, the assumption of neutrality or uniform ionisation is not applicable. As

such, an alternative model revising the relevant assumptions of the published approaches was used to predict the intracellular concentrations in 2D monolayer test systems. Based on partition coefficients between different mediums, and physical volumes, we can predict an apparent volume of distristubtion in the *in vitro* system. Given this volume, we can calculate the unbound medium concentrations, $C_{medium,\,u}$.

$$C_{medium,u} = \frac{C_{nominal} \cdot fu_{FBS,dilu} \cdot V_{medium}}{V_{medium} + k_{air} f_{ui} V_{air} + k_{cell} V_{totalcell} + k_{plastic} SA_{medium} \cdot 10^3}$$

Where $C_{nominal}$ is the nominal concentration, V_{medium} (L) is the volume of culture medium, V_{air} (L) the volume of air in the headspace above the media, and $V_{totalCell}$ (L) is the total volume of cultured cells at the time of the assay, SA_{medium} (m^2) is the surface area of plastic in direct contact with culture medium, $fu_{FBS,dilu}$ is the fraction unbound in fetal bovine serum accounting for the dilution of FBS in culture (where FBS is not included in the culture medium $fu_{FBS,dilu}$ =1), the partition coefficients between culture medium and air, cells and plastic are k_{air}, k_{cell}, and $k_{plastic}$, respectively, and are defined below. The fraction unionised, f_{ui}, is calculated based on the Henderson Hasselbalch equation using the compound specific pKa and the compartment relevant pH.

$$f_{ui} = \frac{1}{1 + Y}$$

$$Y_{neutral} = 0$$

$$Y_{acid} = 10^{(pH - pKa)}$$

Binding to Serum Components

The predominant binding protein present in untreated fetal bovine serum (FBS) is albumin. FBS also contains lipids and free fatty acids. The lipids within FBS are diverse and not individually characterised or quantified routinely. However, the neutral lipid triacylglyceride (TAG) is routinely quantified and reported in the certificate of analysis, as is the concentration of albumin. Assuming albumin and TAG to be the most significant binding components in FBS and complete cell culture medium, we can predict the fraction unbound in FBS, fu_{FBS}.

$$fu_{FBS} = \frac{1}{1 + K_{protein} f_{protein} + \frac{P_{nl} f_{nl,FBS}}{1 + Y_{FBS}}}$$

Where $f_{protein}$ (v/v) is the fraction of FBS comprised of protein, $K_{protein}$ is the albumin:water partition coefficient, P_{nl} is the neutral lipid partition coefficient (defined below), and f_{nl} is the fraction of FBS comprised of neutral lipid. Assuming that the fraction of albumin in the FBS is representative of the total protein fraction responsible for protein binding in the FBS, this can be calculated from the mass of albumin reported in the certificate of analysis for a batch of FBS.

$$f_{protein} \approx f_{alb,FBS} = \frac{mass\ albumin \cdot PSV_{al\ bumin}}{1000}$$

Where $PSV_{albumi}\ n$ is the partial specific volume of albumin (0.73 mL/g (Kupke, Hodgins *et al.* 1972)). The albumin to water partition coefficient, $K_{protein}$, can be determined experimentally or can be calculated based on a previously described relationship with the octanol to water partition coefficient (Endo and Goss 2011).

if $logP_{ow} < 4.5$

$$\log k_{albumin} = 1.08 \cdot \log P_{ow} - 0.7$$

if $logP_{ow} \geq 4.5$

$$\log k_{albumin} = 0.37 \cdot \log P_{ow} + 2.56$$

In much the same way the volumetric fraction of neutral lipid in FBS can be calculated, using TAG as a surrogate for neutral lipid content. TAG concentration is routinely determined using an enzymatic assay and so reported as a molar concentration.

$$f_{nl,FBS} \approx f_{TAG} = \frac{[TAG] \cdot 10^{-3} \cdot MW_{TAG} \cdot PSV_{TAG}}{1000}$$

Where PSV_{TAG} is the partial specific volume of TAG (1.09 mL/g (Deckelbaum, Granot *et al.* 1984)) and the molecular weight of TAG is taken to be 885.453 g/mol; specifically, this corresponds to the molecular weight of trioleate, a TAG molecule comprising a glycerol backbone and three oleic acid residues.

A dilution factor, D, can then be calculated to correct fu_{FBS} for the volumetric fraction of media comprising serum, f_{serum}, and so $fu_{FBS,dilu}$ can be calculated.

$$D = \frac{1}{f_{serum}}$$

$$fu_{FBS,dilu} = \frac{fu_{FBS}}{\frac{1}{D} \cdot (1 - fu_{FBS}) + fu_{FBS}}$$

Calculation of Partition Coefficients

The partition coefficient between the culture medium and air is derived from the test compounds Henry's Law constant which may be determined experimentally or predicted using a variety of *in silico* tools. The dimensionless air medium partition coefficient is determined:

$$k_{air,u} = \frac{k_H}{RT}$$

Where K_H is the Henry's Law constant expressed in SI units (Pa m^3 mol^{-1}), R is the universal gas constant (8.314 Pa m^3 K mol^{-1}), and T is the reference temeprature at which Henry's Law constant as been defined (K).

The plastic to medium partition coefficient is predicted based on the octanol to water partition coefficient using a linear relationship established by Kramer (Kramer 2010) and used in the model published by Comenges *et al* (Zaldivar Comenges, Joossens *et al.* 2017). Note that this partitioncoefficient is not dimensionless and has units (m).

The partition coefficient between cells and culture medium is based on an adaption of the published approach of Rodgers and Rowland for predicting the partitioning of different compound classes between plamsa and tissues, based on composition (Rodgers, Leahy *et al.* 2005, Rodgers and Rowland 2007). We derive $K_{cell\ uu,uu}$ and $K_{iw\ uu,uu,\ organelle}$ from the steady-state Fick-Nernst-Planck equation to describe the passive permeation of electrolytes across the cell and organelle membranes, respectively, as well as the passive permeation of neutral molecules,

$$k_{cell\ uu,uu} = \frac{1 + \frac{P^{unbound,ionised}}{P^{unbound,unionised}} \frac{N}{e^{N_1} - 1} Y_{ew}}{1 + \frac{P^{unbound,ionised}}{P^{unbound,unionised}} \frac{N}{e^{N_1} - 1} e^{N} Y_{iw}}$$

$$k_{IW_{organelle}^{uu,uu}} = \frac{1 + \frac{P^{unbound,ionised_1}}{P^{unbound,unionised}} \frac{N_1}{e^{N_1} - 1} Y_{iw}}{1 + \frac{P^{unbound,ionised_1}}{P^{unbound,unionised}} \frac{N_1}{e^{N_1} - 1} e^{N_1} Y_{organelle}}$$

$$N_{neutral} = 0$$

$$N_{monoacid} = -\frac{\emptyset F}{RT}$$

Where P is the permeability coefficient of either the ionised or unionised moiety. Using a ratio of the permeability coefficient between the unionised and ionised species we can describe the differential permeability of the two molecular forms. It has previously been assumed that the permeability coefficient of the ionised species is 3-4 log units lower than that of the neutral form (Trapp, Rosania *et al.* 2008). Here we assume ionised, unionised permeability coefficient ratio of 3.3 log units for the monoprotic anion. F is the Faraday constant (96484.56 C mol^{-1}) and ø is the cell membrane potential (V).

$$k_{cell,u} = \frac{C_{cell}}{C_{media_{unbound}}}$$

$$= \begin{pmatrix} (1 - f_{lyso} + f_{mito})(f_{iw}(1 + Y_{iw}) + P_{nl} f_{nl} + P_{np} f_{np}) \\ + f_{lyso}(f_{iw}(1 + Y_{lyso}) + P_{nl} f_{nl} + P_{np} f_{np}) K_{iw_{lyso}^{uu,uu}} \\ + f_{mito}(f_{iw}(1 + Y_{mito}) + P_{nl} f_{nl} + P_{np} f_{np}) K_{iw_{mito}^{uu,uu}} \end{pmatrix} \frac{1}{1 + Y_{ew}} k_{cell\ uu,uu}$$

Incorporating $K_{cell\ uu,uu}$ and $K_{iw\ uu,uu,\ organelle}$ we adapt the original Rodgers and Rowland approach where f_{iw}, f_{lyso}, f_{mito}, f_{nl}, f_{np} denote the fractional cellular volumes of the intracellular water, lysosomes, mitochondria, neutral lipids, and neutral phospholipids, respectively. P_{nl} and P_{np} describe the partitioning of the compound between intracellular water and neutral lipids and neutral phospholipids, respectively. Where the olive oil to water ($P_{vo:w}$) and octanol to water ($P_{o:w}$) partition coefficients are used as surrogates, respectively.

This revises a fundamental assumption of the published Rodgers and Rowland approach, and previously published biokinetic models, that only unionised molecular species can passively traverse biological membranes (Rodgers, Leahy *et al.* 2005, Rodgers and Rowland 2007, Armitage, Wania *et al.* 2014, Zaldivar Comenges, Joossens *et al.* 2017). The approach also expands on previously published biokinetic models by describing distribution into two subcellular organelle compartments (lysosome and mitochondria). Based on the approach described above, total intracellular concentrations corresponding to nominal effect concentrations determined experimentally can then be calculated.

$$C_{cell} = k_{cell,u} \cdot C_{media,dissolved,u}$$

It should be noted that while this approach can be used to model ampholytes, monoprotic and diprotic acids and bases, the equations above are described in forms with specific to neutral compounds and monoprotic acids, relevant to the compounds investigated in this read acros. A describition of this model has been presented previously (Fisher, Jamei *et al.* 2017), and is currently submitted for peer-review.

Summary of Model Assumptions

The model described above, like those previosuly published (Armitage, Wania *et al.* 2014, Fischer, Henneberger *et al.* 2017), is a steady-state approximation of multiple dynamic processes. Critically, it assumes a closed system such that the loss of test compounds within the *in vitro* system is assumed to be negligible with no metabolic clearance or instability. When that is not the case, the steady-state assumption may lead to overestimation of intracellular concentrations of the parent molecule, particularly for highly metabolised chemicals. In line with the assumption of a closed system, the culture system is assumed to be hemetically sealed, such that the air above the culture medium is a defined volume. For volatile compounds, tested in unsealed systems, this could also result in an overprediction of intracellular compounds, since distribution into the air will act as a clearance mechanism. As part of the steady-state assumption the model assumes that the volume of cultured cells is constant with no significant increase or decrease over the course of the test assay. Finally the model assumes that all binding and partitioning processes are non-saturable, with the distribution of test compound into the cell mediated through passive diffusion. The model assumes that there are no active uptake or efflux tranpsort processes relevant to the partitioning of the test compound within the test system.

Alternative Biokinetic Predictions

An assumption of the biokinetic model used to predict intracellular concentrations is that cells are cultured as a 2D monolayer. Permeability into cells cultured in multi -layer, three dimensional systems may not be well described based on the approach detailed above; particularly permeability into cells that may not be in direct contact with the culture medium (i.e sandwiched between adjacent cells). Thus, we adopt a simplified approach to predict the corresponding concentration of free chemical in test medium, as if the determined effective concentration was made up in complete culture medium prior to cell treatment. This predicted unbound effective concentration can then be translated to an unbound plasma concentration *in vivo*; such approaches have been described previously (Gulden and Seibert 2003).

As described above, the predominant binding components present in untreated FBS are albumin and neutral lipids. In order to predict the free concentration of test compounds in treatment medium it is necessary to account for binding to these media components. Here we assume that the binding to albumin and lipid (TAG) in complete culture (treatment) medium, are the only significant processes limiting the availability of test compound for distribution into subsequently treated cells. Thus, loss of compound due to volatility, or the binding to the plastics used in cell culture are not accounted for in this approach.

If we consider the threshold for significant volatility to be an air-water partition coefficient (K_{air} <0.03), as previously assumed by Fischer *et al*[1], then the assumption that volatility has no significant impact on the freely dissolved media concentration holds for all of the compounds investigated here. Polymer-water partition coefficients have been shown to be significantly lower than octanolwater partition coefficients (P_{ow})(Armitage, Wania *et al.* 2014), here used to determine the partitioning to TAG and albumin, as described below. As such, we assume here that binding to plastics used in the handling and preparation of culture medium have no significant impact on the free concentration of test compound. Finally, assuming that the maximal tested concentration does not exceed the solubility of the compound in complete culture medium, and taking the binding to protein and lipid in culture media to be linear across the tested concentration range, we can calculate an

unbound fraction of test compound given the composition of the medium and the P_{ow} of the test compound.

$$fu_{media} = \frac{1}{1 + K_{albumin} f_{albumin} + \frac{P_{vow} f_{nl}}{1 + Y}}$$

$$Y_{neutral} = 0$$

$$Y_{monoprotic\ acid} = 10^{(pH - pKa)}$$

$$\log K_{vow} = 1.115 \cdot \log K_{OW} - 1.35^{4}$$

Where $f_{albumin}$ is the volumetric fraction of medium comprised of protein, $K_{albumin}$ is the albumin-water partition coefficient, K_{vow} is the olive oil-water partition coefficient (derived from P_{ow}), fnl is the volumetric fraction of medium comprised of neutral lipid (TAG), and Y is the ratio of the ionised to unionised concentrations in the culture medium calculated using the Henderson-Hasselbalch equation. As above, the binding of test compound to neutral lipids is assumed to be limited to the unionised fraction of the solubilised compound in medium. The volumetric fractions of albumin and TAG in the culture medium can calculated using an analogous approach to that described above using the part ial specific volumes of albumin and TAG.

$$f_{albumin} = \frac{[albumin] \cdot PSV_{albumin}}{1000}$$

$$f_{nl} \approx f_{TAG} = \frac{[TAG] \cdot PSV_{TAG}}{1000}$$

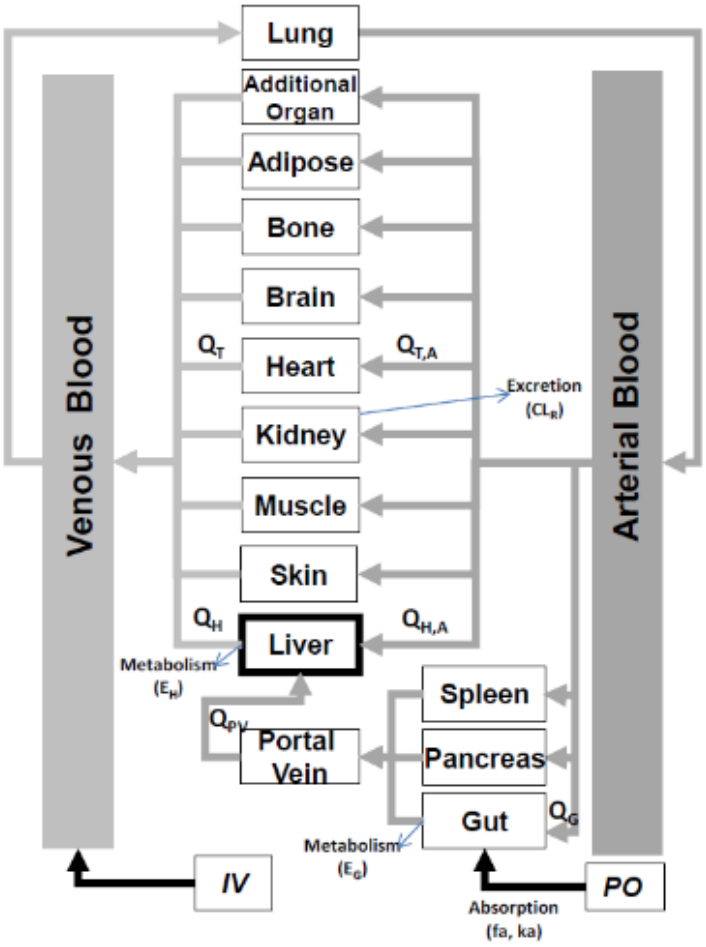
It should be noted that this relationship assumes that the neutral and the ionised fractions of compound partition equally into the hydrophobic phase and so $logK_{albumin}$ is not influenced by the ionisation state for ionisable compounds. Here, both the concentration of albumin and TAG are taken to be in units of mg/mL.

4. PBPK modelling

PBPK models for the compounds were constructed in the rat and human Simcyp Simulator (V17r1, Certara Ltd. Simcyp Division, Sheffield, UK; www.simcyp.com). The Simcyp simulator has been extensively tested and used by a consortium of industry, regulatory and academic scientists. Lists of known bugs within the code of the Simcyp simulator are maintained on our website (https://members.simcyp.com/account/softwareIssues/). The QA system used to support the production of each version of the Simcyp simulator has been described in detail (Jamei, Marciniak *et al.* 2013).

Physiologically based pharmacokinetic models

An aim of the read-across case study was to estimate the concentrations of the compounds in different tissues of the body specifically the liver which is the target organ of interest for these agents in the steatosis AOP under investigation. To accomplish this a full body physiologically based pharmacokinetic (PBPK) model was used; a schematic framework for this model is shown below. In the human simulator, the ability to add further specific organs as an additional organ is available, but in the simulations for this read-across study, this functionality was not required.



In the PBPK schematic above Q_H, Q_{HA}, Q_{PV}, Q_G, $Q_{T,A}$ and Q_T are blood flows in the total hepatic, hepatic artery, hepatic portal vein, gut and blood flows into and out of the other tissue (T) compartments, respectively; E_G and E_H are the fractions undergoing first pass metabolism in the gut and liver, respectively; CL_R is the renal clearance; fa and ka are the fraction absorbed and the first order absorption rate constant, respectively.

Distribution

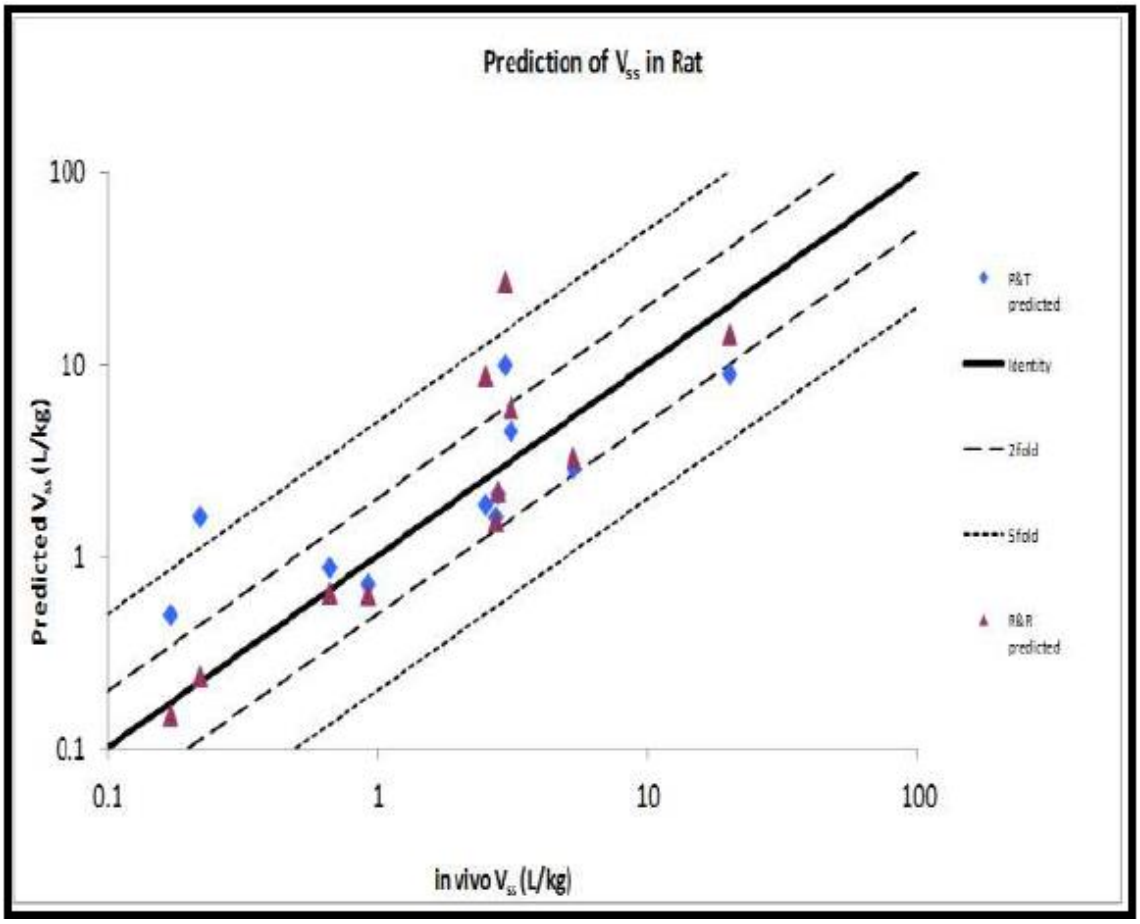
In the human simulator, inter-individual variability in tissue distribution is accounted for through relationships between tissue volume and age, sex, weight and height (Jamei, Dickinson *et al.* 2009). In both the human and rat PBPK models the *in vivo* volume of distribution at steady state (V_{ss}) is predicted using Equation 1 from Sawada *et al.* (Sawada, Hanano *et al.* 1984).

$$V_{ss} = V_p + V_e \cdot E{:}P + \sum V_t \cdot K_{p,t}$$

Equation 1

Where V is the fractional body volume (L/kg) of a tissue (t), erythrocytes (e), and plasma (p), E:P is the erythrocyte: plasma drug concentration ratio and $K_{p,t}$ is the partition coefficient of drug between tissues and plasma components. Three methods are available for prediction of $K_{p,t}$. The first method was reported by Poulin and Thiel (Poulin and Theil 2002) as corrected by Berezhkovskiy (Berezhkovskiy 2004) (Method 1). Method 1 uses the physicochemical properties of the compounds (pKa and log P) together with *in vitro* information (B/P and fu) to predict partitioning into the tissues with the assumption that tissues and plasma are mixtures of lipids, water and proteins with a global pH of 7.4. The second method was developed by Rodgers and Rowland (Rodgers and Rowland 2006) (Method 2). The latter splits the tissue water volume into intra- and extracellular components, with the addition of an acidic phospholipid fraction within tissues. These equations take explicit account of the extent of ionisation of a compound at the pH of the compartment concerned and have been shown to improve the prediction of tissue:plasma partition coefficients, and consequently V_{ss}, for strong bases. The Rodgers and Rowland method was further extended by the science team at Simcyp to account for ion permeability and the effect of differences in membrane potential in the different tissues on compound distribution (Method 3); however, Method 3 is currently only available in the human simulator. Method 3 also allows for the distribution of compounds into specific subcellular organelles to be modelled and forms the basis for the biokinetic model developed within the EUTOXRISK project (Fisher, Gardner *et al.* 2017). These mechanistic predictions assume non-saturating conditions prevail for all binding processes, drug transport is via passive processes (i.e. no active transport), and each tissue has a well –stirred distribution limited by blood perfusion (i.e. tissues are considered as perfusion limited not permeability limited).

Performance verification for both Method 1 and Method 2 in the rat with a wide range of compounds (molecular weight of 192-1202 kDa, 19% acidic, 46% basic, 9% neutral, 27% ampholyte) is shown in the figure below. The range of *in vivo* V_{ss} for the studied compounds was 0.17-19.9 L/kg and for Method 1 and Method 2, 64% and 82% of predictions were within 2-fold of observed values, respectively.



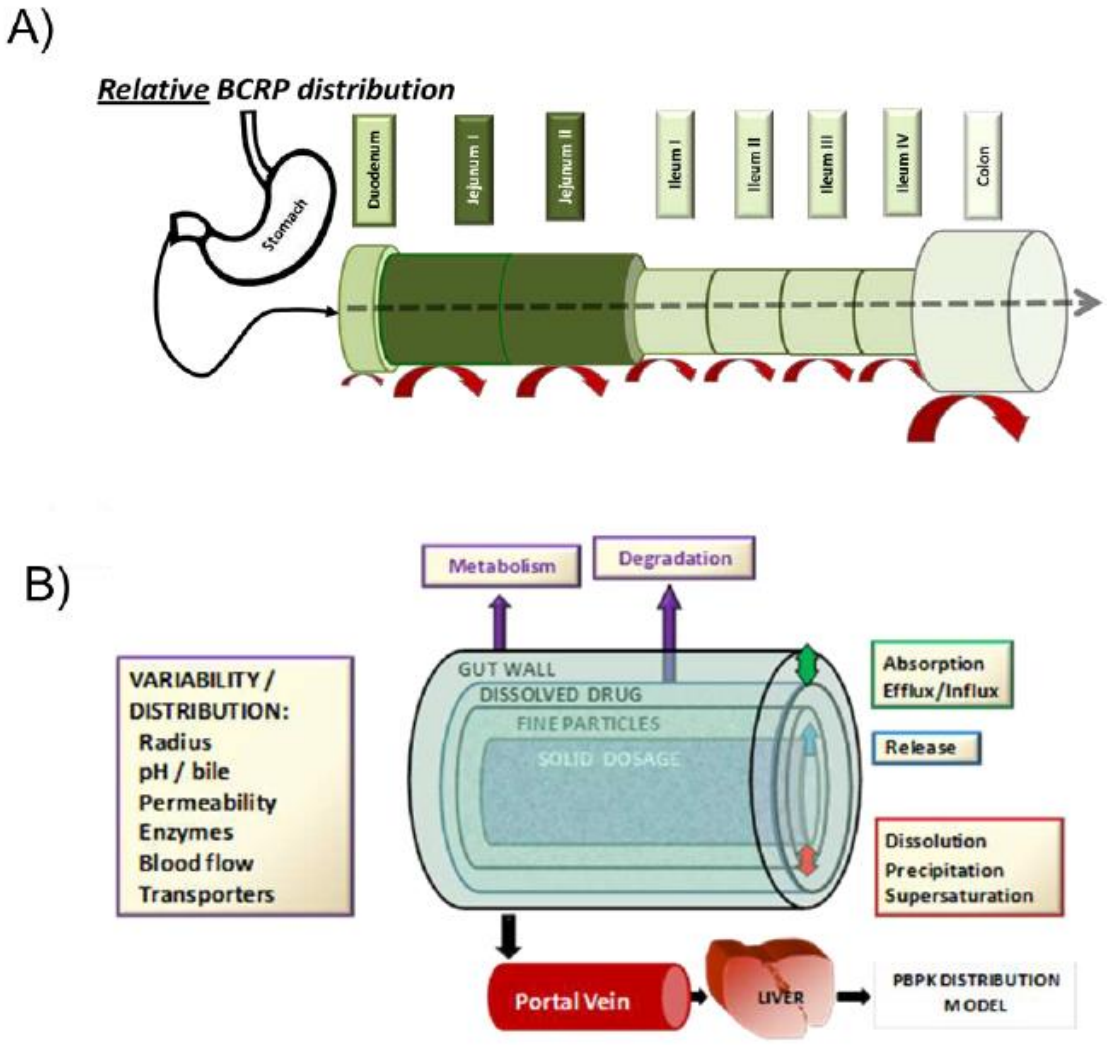
Perfusion and permeability limited distribution

Lipophilic drugs diffuse rapidly across the capillary membrane into tissue interstitial fluid such that blood flow to the tissue is the rate-limiting step in uptake. This is described as perfusion-limited distribution and is implemented in all tissues represented in the PBPK model. In addition, an option is provided within the Simcyp Simulator to allow for permeability-limited uptake, simulating both passive diffusion in parallel with active uptake and efflux in specific organs such as the liver, kidney (human only), intestine and brain. In these models, the tissue is divided into compartments representing vascular, extracellular and intracellular fluid spaces - with distribution between these spaces defined as a dynamic process. The 'permeability-limited' models in the liver, brain and kidney are only available when Method 2 or 3 are selected to predict $K_{p,t}$. Given the lipophilic nature of the compounds in the read- across case study, perfusion limited models were used for all tissues in the simulations run in both rat and human in this read-across study.

Oral Absorption

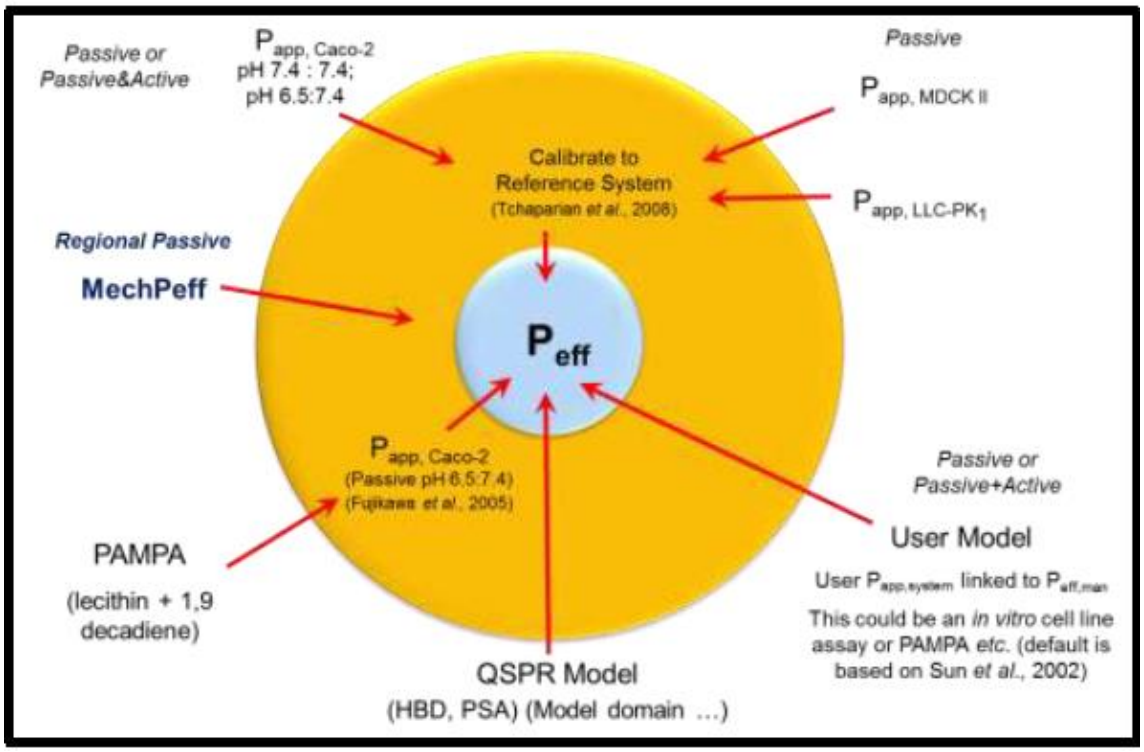
For drugs in solution, several absorption models are available within the Simcyp Simulator including a first-order absorption model, a compartmental absorption transit (CAT) model (Yu and Amidon 1998) and the advanced dissolution absorption metabolism (ADAM) model (Jamei, Turner *et al.* 2009). Simulation of the absorption of compounds from solid dosage forms requires use of the ADAM model. The ADAM model, as implemented in the

Simcyp Simulator, divides the gastrointestinal tract (GIT) into nine anatomically defined segments from the stomach through the intestine to the colon. Drug absorption from each segment is described as a function of release from the formulation, dissolution, precipitation, luminal degradation, permeability, metabolism, transport and transit from one segment to another. It is assumed that absorption from the stomach is insignificant compared with that from the small intestine, and that movement of liquid and solid drug through each segment of the GIT may be described by first -order kinetics. Dissolution rate from solid dosage forms is calculated from information on drug aqueous solubility and particle size using diffusion layer models (DLM) (Wang and Flanagan 1999, Wang and Flanagan 2002).



The diagrames above show the structure of the ADAM model in which the GI tract is divided into 9 sections with segregated blood flows to each section. The abundance of various enzymes and transporters in each segment varies non-monotonically along the intestine as indicated by the varying intensity of the colour for each section (BCRP distribution is indicated in A) (Harwood, Neuhoff *et al.* 2013, Harwood, Achour *et al.* 2016, Harwood, Achour *et al.* 2016). B shows segments of the small intestine indicating the various processes that can be simulated (from (Darwich, Neuhoff *et al.* 2010)).

The effective permeability in humans, $P_{eff, man}$ (jejunal), can be measured using the Loc-i-gut methodology and can be used in simulations to describe the absorptive processes in the intestine (Nilsson, Fagerholm *et al.* 1994). For novel investigation drugs, many marketed agents, and all industrial chemicals, measured values of $P_{eff, man}$ (jejunal) are not available and therefore several methods can be used within the Simcyp Simulator to predict $P_{eff, man}$ (jejunal). These are based on data obtained with cell lines (such as Caco-2, MDCK-II or LLC-PK1 cells)(Sun, Lennernas *et al.* 2002), PAMPA or from a QSPR model based upon physicochemical properties (PSA and HBD,(Winiwarter, Bonham *et al.* 1998)) or by using the mechanistic permeability (MechPeff) model (Fehler! Verweisquelle konnte nicht gefunden werden.). The regional permeability (seven small intestine segments plus colon) for all of the methods (apart from MechPeff) is assumed to be the same by default but can be modified by the user. The regional distributions of drug metabolising enzymes and efflux transporters such as P-gp and BCRP are also incorporated, allowing simulation of the effects of efflux transport and metabolism on drug absorption.



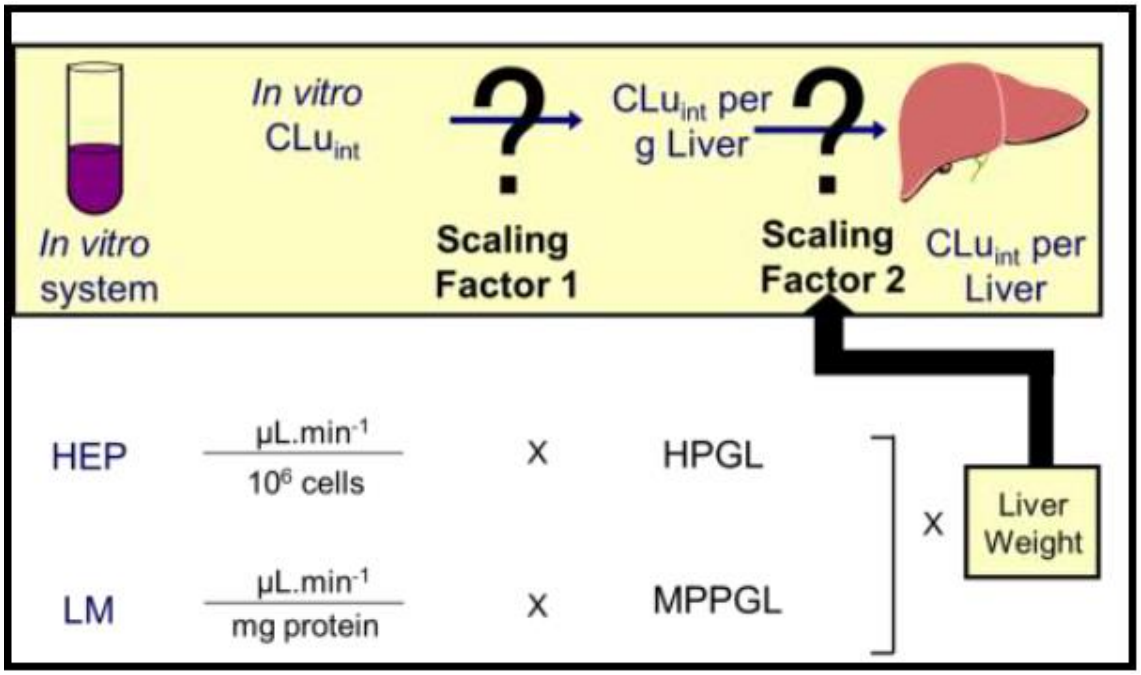
Input options to predict Peff and subsequently absorption within the Simcyp Simulator are shown in the diagram above. The physicochemical based QSPR (HBD and PSA) model was used in the simulations in this study. In human simuations a simple first-order absorption model used, while in rat simulations of VPA incorporating enterohepatic recirculation, the ADAM model was used predicting Peff from HBD and PSA.

Metabolic Clearance

Elimination of a compound can be characterised by various inputs of clearance such as intravenous or oral clearance (CL_{iv} or CL_{po}), whole organ metabolic clearance via hepatocytes (CL_{int}; µL/min/10^6 cells), liver and intestinal microsomes (CL_{int}; µL/min/mg microsomal protein) or incubation with intestinal slices (CL_{int}; µL/min/g of intestine).

Hepatic metabolic clearance

On a general basis the *in vivo* hepatic metabolic clearance is predicted using *in vitro-in vivo* extrapolation (IVIVE), as shown schematically below, followed by scaling for the specific metabolising tissue blood flow (liver in this case) and fraction of unbound drug in blood.



For the human simulator clearance was predicted using metabolic intrinsic clearance ($CL_{int, hep}$) data generated in human hepatocytes. *In vitro* $CL_{int, hep}$ was scaled up to the *in vivo* CL_{int} ($CL_{int, u, H}$) according to Equation 2.

$$CL_{int\ u,H} = \frac{CL_{int,hep}}{fu_{hep}} \cdot uptake \cdot HPGL \cdot LW \cdot 10^{-6} \cdot 60$$

Equation 2

Where fu_{hep} is the fraction unbound of the compound in the hepatocyte incubation, HPGL is the number of hepatocytes per gram of liver, uptake was assumed to be only be due to passive processes in these simulations (uptake =1), and 10^{-6} and 60 are to adjust units from µl/min/10^6 cells to L/h in the whole liver. Correction for non-specific protein binding is important for IVIVE (McGinnity, Berry *et al.* 2006, Brown, Chadwick *et al.* 2007) and $fu_{,hep}$ was predicted as described by Kilford *et al* (Kilford, Gertz *et al.* 2008), Equation 3.

$$fu_{hep} = \frac{1}{1 + \frac{K_{HM}}{K_{mic}} \cdot V_R \cdot 10^{0.072 \cdot \log D_{7.4}{}^2 + 0.067 \cdot \log D_{7.4} - 1.126}}$$

Equation 3

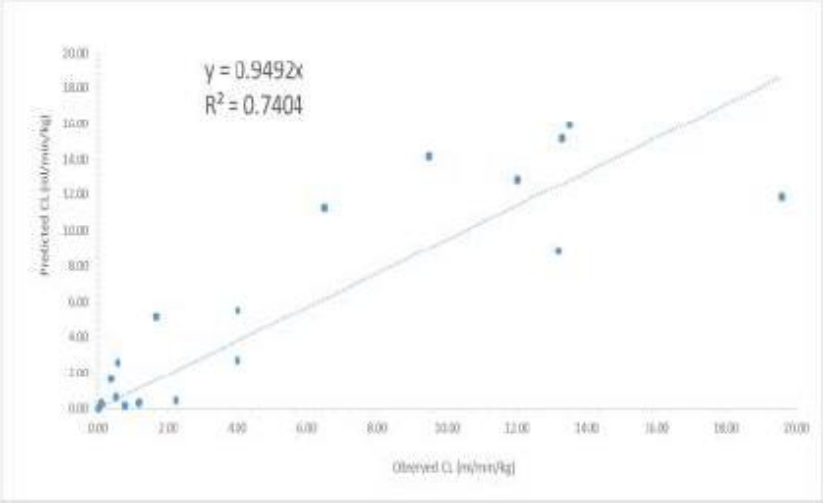
Where V_R is a V_{cell}/V_{inc} ratio, where V_{cell} is the cell volume and V_{inc} the incubation volume. A K_{HM}/K_{mic} ratio of 125 was assumed (Kilford, Gertz *et al.* 2008). V_R is 0.005 at the cell concentration of 10^6 cells/ml (Kilford, Gertz *et al.* 2008), and was normalised for P of 1 mg/ml.

The hepatic clearance (CL_H) in humans was calculated from the whole liver scaled *in vivo* $CLint_{u,h}$ using the well stirred model Equation *4*.

$$CL_{H,b} = \frac{Q_H \cdot Clint_{u,h} \cdot fu_B}{Q_H + (Clint_{u,h} \cdot fu_B)}$$

Equation 4

Where Q_H = hepatic blood flow, $CLint_{u,h}$ is the *in vivo* hepatic intrinsic clearance and fu_B = fu/(B/P). The accuracy of this *in vitro – in vivo* approach to predict human clearance from data generated in human hepatocytes; the relationship between predicted and observed human clearance for a series of 18 compounds is shown below. The predicted clearance was made using IVIVE approaches described above based on the *in vitro* intrinsic clearance generated in human hepatocyte incubations previously performed by Cyprotex. The range of LogP values for these compounds was -0.07 to 4.8, the target and all source compounds in this read-across fall within this range.



Clearance inputs in the rat models

Equivalent *in vitro* metabolic data in rat hepatocytes could not be generated as part of the EUTOX-RISK project. A simple allometric scaling approach was liable to result in an under-estimation of VPA exposure in rat, due to the role of EHR in rat VPA kinetics. Therefore, a reverse translation approach was adopted, back-calculating intrinsic hepatic clearance from published *in vivo* clearance (Kameya, Hokama *et al.* 2009) data using the below equations. Simply, this calculates the whole liver clearance and then uses the reverse well-stirred liver model to determine heaptic intrinisc clearance.

$$CL_{H,b} = \frac{CL_{iv} - CL_R - CL_{add}}{BP}$$

$$CL_{int,u,h} = \frac{Q_H \cdot CL_{H,b}}{fu_b \cdot (Q_H - CL_{H,b})}$$

$$fu_{u,b} = \frac{fu}{BP}$$

Enterphepatic recirculation in the rat PBPK model

In rat, metabolite cleared through the biliary route is deconjugated in the gut, thus back-converted to the parent compound and so available for reabsoprtion from the intestine. The concentration of metabolite (X) available for back-conversion is described the following ODE:

$$\frac{dX}{dt} = f \cdot CL_{bile} \cdot uptake \cdot \frac{fu}{BP} \cdot \frac{C_{liver}}{\frac{K_{p,liver}}{BP}} - K_{bc} \cdot X$$

Where CL_{bile} is the biliary clearance rate, uptake is an empirical correction for active uptake into the liver (equals one in rat VPA simualtions), K_{bc} is the rate constant for the back-conversion of metabolite to parent compound, and f is the fraction of biliary metabolite avauiolable for back -conversion. Back-converted metabolite was added to the duodenal compartment of the ADAM model used in rat simualtions consistent with physiological bile flow into the dudodenum.

Excretion

It is possible to account for the excretion of unchanged drug in the kidney or via the biliary system in the PBPK models, but in human the models developed here, hepatic metabolism was considered as the only route of clearance. Bililary clearance was only considered for the metabolites of VPA in the rat PBPK model incorporating EHR.

General description of equations for eliminating and non-eliminating organs

Each compartment within the full body PBPK model is initially described as a perfusion limited model with representative equations for an eliminating and non-eliminating organ as shown below. The equations describing the behaviour of the compound within the intestine following oral absorption are as described by Jamei *et al* (Jamei, Turner *et al.* 2009). The basic principles for a PBPK model outlined on P19 of the WHO guidance on PBPK modelling were adhered to, namely:

1) the mixing of the chemical in the effluent blood from the tissues is instantaneous and complete;
2) blood flow is unidirectional, constant and non-pulsatile; and
3) the presence of chemicals in the blood does not alter the blood flow rate

In non-eliminating tissues the compound concentration (C) at a given time (t) is defined by Equation 5. Where Q = blood flow to the tissue, V = tissue volume, C_{ab} = arterial blood concentration, BP = blood to plasma ratio and $K_{p,\,t}$ is the partition coefficient of drug between tissue (t) and plasma. In the liver, the Equation 6 is applied.

$$\frac{dC_{tissue}}{dt} = \frac{Q_{tissue}}{V_{tissue}}\left(C_{ab} - \frac{C_{tissue}}{K_{p,t}/_{BP}}\right)$$

Equation 5

$$\frac{dC_{liver}}{dt} = \frac{1}{V_{liver}}\left((Q_{liver} - Q_{pv})C_{ab} + Q_{pv}C_{pv} - \frac{Q_{liver}C_{liver}}{K_{p,t}/_{BP}} - \frac{f_u}{BP} \cdot \frac{CLu_{int}}{K_{p,t}/_{BP}} \cdot C_{liver}\right)$$

Equation 6

Where Q_{liver} = sum of blood flow to the liver by the hepatic artery and hepatic portal vein, Q_{pv} = hepatic portal blood vein flow, C_{pv} = hepatic portal compound vein concentration, fu = fraction unbound in plasma, CLu_{int} = intrinsic clearance

Within each simulated animal or human subject the sum of the tissue blood flow rates (excluding the lung) are equal to cardiac output. In line with accepted mammalian physiology, the lung receives a blood flow equal to total cardiac output. Tissue volumes and blood flow rates are within the documented range for each species considered (Jamei, Dickinson *et al.* 2009, Musther, Harwood *et al.* 2017).

Population data

Predictions of plasma drug exposure, clearance and other parameters such as fraction metabolised by a particular pathway were made for virtual populations of healthy volunteers. Each population is generated using values and formulae describing demographic, anatomical and physiological variables. Thus, in order to assess clearance predictions in a specific population, data are required for the population variables as well as for the *in vitro* metabolism/transport of the test drug and its observed clearance in the population of interest. The parameter values within the Simcyp Simulator for creating a virtual healthy volunteer population (population, physiological parameters including liver volume and blood flows, enzyme abundances) have been described previously (Jamei, Dickinson *et al.* 2009).

Physiology data used in the healthy human population PBPK populations

A virtual population of 100 individuals aged 20-50 with 50% of the subjects being female was used for the simulations, the range of physiology data used in the human population PBPK simulations is summarised below.

Parameter	Mean value	Range
Age (y)	29.6	20 – 48
Weight (kg)	72.45	45-118
Height (cm)	168.4	149 – 189
Cardiac output (L/h)	321.8	252 – 416
Serum albumin (g/L)	45.7	38 - 60

Data for simulation of the human *in vivo* kinetics of read-across study compounds

Data and the corresponding source and/or reference used in the compound files are shown in the tables below.

Input parameter values used to simulate the kinetics of Valproic acid (99-66-1)

Parameter	Value	Method / Comment	Source/Reference
MW [g/mol]	144.21		EPI-Suite (v4.1, US-EPA)
logP	2.75	experimental	EPI-Suite (v4.1, US-EPA)
Compound Type	Monoprotic acid		
pKa	4.8		ACD/Percepta (2012 release, build 2254, ACD Labs)
TPSA (Å^2)	37.3		https://pubchem.ncbi.nlm.nih.gov/compound/3121#section=Chemical-and-Physical-Properties
Hydrogen bond donors	1		https://pubchem.ncbi.nlm.nih.gov/compound/3121#section=Chemical-and-Physical-Properties
fu	0.138 0.310	predicted predicted	Lhasa CORAL model
B/P ratio	0.55	assumed	
fa	0.996	predicted using HBD and PSA (Simcyp v17r1)	(Winiwarter, Bonham *et al.* 1998, Winiwarter, Ax *et al.* 2003)
ka (h^{-1})	2.546	predicted (Simcyp V17r1)	
fu_{Gut}	1	assumed	
CL_{int} (µl/min/10^6 cells)	0.219	experimental (HµREL co-culture system)	Cyprotex data (CYP1440)
Hepatocyte binding (fu_{heps})	0.954	predicted	(Kilford, Gertz *et al.* 2008)

Input parameter values used to simulate the kinetics of 2-Ethyl Butyric acid (88-09-5)

Parameter	Value	Method / Comment	Source/Reference
MW [g/mol]	116.16		EPI-Suite (v4.1, US-EPA)
logP	1.68	experimental	EPI-Suite (v4.1, US-EPA)
Compound Type	Monoprotic acid		
pKa	4.8		ACD/Percepta (2012 release, build 2254, ACD Labs)
TPSA (Å^2)	37.3		https://pubchem.ncbi.nlm.nih.gov/compound/6915#section=Chemical-and-Physical-Properties
Hydrogen bond donors	1		https://pubchem.ncbi.nlm.nih.gov/compound/6915#section=Chemical-and-Physical-Properties
fu	0.348 0.712	predicted predicted	Random forest (model 1) Lhasa
B/P ratio	0.55	assumed	
fa	0.996	predicted using HBD and PSA (Simcyp v17r1)	(Winiwarter, Bonham *et al.* 1998, Winiwarter, Ax *et al.* 2003)
ka (h^{-1})	2.546	predicted (Simcyp V17r1)	
fu_{Gut}	1	assumed	
CL_{int} (µl/min/10^6 cells)	9.62	experimental (HµREL co-culture system)	Cyprotex data (CYP1440)
Hepatocyte binding (fu_{heps})	0.955	predicted	(Kilford, Gertz *et al.* 2008)

Input parameter values used to simulate the kinetics of 2-Ethyl Hexanoic acid (149-57-5)

Parameter	Value	Method / Comment	Source/Reference
MW [g/mol]	144.21		EPI-Suite (v4.1, US-EPA)
logP	2.64	experimental	EPI-Suite (v4.1, US-EPA)
Compound Type	Monoprotic acid		
pKa	4.8		ACD/Percepta (2012 release, build 2254, ACD Labs)
TPSA (Å^2)	37.3		https://pubchem.ncbi.nlm.nih.gov/compound/8697#section=Chemical-and-Physical-Properties
Hydrogen bond donors	1		https://pubchem.ncbi.nlm.nih.gov/compound/8697#section=Chemical-and-Physical-Properties
fu	0.142 0.310	predicted predicted	Random forest (model 1) CORAL model
B/P ratio	0.55	assumed	
fa	0.996	predicted using HBD and PSA (Simcyp v17r1)	(Winiwarter, Bonham *et al.* 1998, Winiwarter, Ax *et al.* 2003)
ka (h^{-1})	2.546	predicted (Simcyp V17r1)	
fu_{Gut}	1	assumed	
CL_{int} (µl/min/10^6 cells)	0.551	experimental (HµREL co-culture system)	Cyprotex data (CYP1440)
Hepatocyte binding (fu_{heps})	0.950	predicted	(Kilford, Gertz *et al.* 2008)

Input parameter values used to simulate the kinetics of 2-Ethyl Pentanoic acid (20225-24-5)

Parameter	Value	Method / Comment	Source/Reference
MW [g/mol]	130.19		EPI-Suite (v4.1, US-EPA)
logP	2.23	experimental	EPI-Suite (v4.1, US-EPA)
Compound Type	Monoprotic acid		
pKa	4.8		ACD/Percepta (2012 release, build 2254, ACD Labs)
TPSA (Å^2)	37.3		https://pubchem.ncbi.nlm.nih.gov/compound/30020#section=Chemical-and-Physical-Properties
Hydrogen bond donors	1		https://pubchem.ncbi.nlm.nih.gov/compound/30020#section=Chemical-and-Physical-Properties
fu	0.296 0.489	predicted predicted	Random forest (model 1) Lhasa
B/P ratio	0.55	assumed	
fa	0.996	predicted using HBD and PSA (Simcyp v17r1)	(Winiwarter, Bonham *et al.* 1998, Winiwarter, Ax *et al.* 2003)
ka (h^{-1})	2.546	predicted (Simcyp V17r1)	
fu_{Gut}	1	assumed	
CL_{int} (µl/min/10^6 cells)	0.779	experimental (HµREL co-culture system)	Cyprotex data (CYP1440)
Hepatocyte binding (fu_{heps})	0.957	predicted	(Kilford, Gertz *et al.* 2008)

Input parameter values used to simulate the kinetics of 2-methylhexanoic acid (4536-23-6)

Parameter	Value	Method / Comment	Source/Reference
MW [g/mol]	130.19		EPI-Suite (v4.1, US-EPA)
logP	2.47	experimental	EPI-Suite (v4.1, US-EPA)
Compound Type	Monoprotic acid		
pKa	4.8		ACD/Percepta (2012 release, build 2254, ACD Labs)
TPSA (Å^2)	37.3		https://pubchem.ncbi.nlm.nih.gov/compound/20653#section=Chemical-and-Physical-Properties
Hydrogen bond donors	1		https://pubchem.ncbi.nlm.nih.gov/compound/20653#section=Chemical-and-Physical-Properties
fu	0.265 0.454	predicted predicted	Random forest (model 3) Lhasa
B/P ratio	0.55	assumed	
fa	0.996	predicted using HBD and PSA (Simcyp v17r1)	(Winiwarter, Bonham *et al.* 1998, Winiwarter, Ax *et al.* 2003)
ka (h^{-1})	2.546	predicted (Simcyp V17r1)	
fu_{Gut}	1	assumed	
CL_{int} (µl/min/10^6 cells)	3.95	experimental (HµREL co-culture system)	Cyprotex data (CYP1440)
Hepatocyte binding (fu_{heps})	0.956	predicted	(Kilford, Gertz *et al.* 2008)

Input parameter values used to simulate the kinetics of 2-methylpentanoic acid (97-61-0)

Parameter	Value	Method / Comment	Source/Reference
MW [g/mol]	116.16		EPI-Suite (v4.1, US-EPA)
logP	1.8	experimental	EPI-Suite (v4.1, US-EPA)
Compound Type	Monoprotic acid		
pKa	4.8		ACD/Percepta (2012 release, build 2254, ACD Labs)
TPSA (Å^2)	37.3		https://pubchem.ncbi.nlm.nih.gov/compound/7341#section=Chemical-and-Physical-Properties
Hydrogen bond donors	1		https://pubchem.ncbi.nlm.nih.gov/compound/7341#section=Chemical-and-Physical-Properties
fu	0.265 0.454	predicted predicted	Random forest (model 1) Lhasa
B/P ratio	0.55	assumed	
fa	0.996	predicted using HBD and PSA (Simcyp v17r1)	(Winiwarter, Bonham *et al.* 1998, Winiwarter, Ax *et al.* 2003)
ka (h^{-1})	2.546	predicted (Simcyp V17r1)	
fu_{Gut}	1	assumed	
CL_{int} (µl/min/10^6 cells)	10.2	experimental (HµREL co-culture system)	Cyprotex data (CYP1440)
Hepatocyte binding (fu_{heps})	0.956	predicted	(Kilford, Gertz *et al.* 2008)

Physiology data used in the rat PBPK simulations

Population variability is not accounted for in the Simcyp rat PBPK simulator and so only population representative simulations were run.

Parameter	Mean value
Weight (kg)	0.25
Cardiac output (ml/min)	78
Serum albumin (g/L)	31.14

Data for simulation of the rat *in vivo* kinetics of Valproic acid

Input parameter values used to simulate the kinetics of Valproic acid in rat (99-66-1)

Parameter	Value	Method / Comment	Source/Reference
MW [g/mol]	144.21		EPI-Suite (v4.1, US-EPA)
logP	2.75	experimental	EPI-Suite (v4.1, US-EPA)
Compound Type	Monoprotic acid		
pKa	4.8		ACD/Percepta (2012 release, build 2254, ACD Labs)
TPSA (Å^2)	37.3		https://pubchem.ncbi.nlm.nih.gov/compound/3121#section=Chemical-and-Physical-Properties
Hydrogen bond donors	1		https://pubchem.ncbi.nlm.nih.gov/compound/3121#section=Chemical-and-Physical-Properties
fu	0.35	experimental	(Loscher 1978)
B/P ratio	0.74	experimental	(Loscher 1978)
$P_{eff,rat}$ (10^{-4} cm/s)	1.65	predicted using HBD and PSA (Simcyp v17r1)	
fu_{Gut}	1	assumed	
CL_{int} (µl/min/mg protein)	10.71	back-calculated from *in vivo* data	(Kameya, Hokama *et al.* 2009)
CL_{int} Bile (metabolite) (µl/min/10^6 cells)	5.8657	back-calculated from *in vivo* data	(Singh, Orr *et al.* 1988)
K_{bc} (h^{-1})	0.08	estimated	
f	0.4	estimated	
CL_R (metabolite) (ml/min)	0.076	experimental	(Singh, Orr *et al.* 1988)

Biokinetic and PBPK References

Armitage, J. M., F. Wania and J. A. Arnot (2014). "Application of mass balance models and the chemical activity concept to facilitate the use of *in vitro* toxicity data for risk assessment." Environ Sci Technol **48**(16): 9770-9779.

Berezhkovskiy, L. M. (2004). "Volume of distribution at steady state for a linear pharmacokinetic system with peripheral elimination." J Pharm Sci **93**(6): 1628-1640.

Blaauboer, B. J. (2010). "Biokinetic modeling and *in vitro-in vivo* extrapolations." J Toxicol Environ Health B Crit Rev **13**(2-4): 242-252.

Brown, H. S., A. Chadwick and J. B. Houston (2007). "Use of isolated hepatocyte preparations for cytochrome P450 inhibition studies: comparison with microsomes for Ki determination." Drug Metab Dispos **35**(11): 2119-2126.

Darwich, A. S., S. Neuhoff, M. Jamei and A. Rostami-Hodjegan (2010). "Interplay of metabolism and transport in determining oral drug absorption and gut wall metabolism: a simulation assessment using the "Advanced Dissolution, Absorption, Metabolism (ADAM)" model." Curr Drug Metab **11**(9): 716-729.

Deckelbaum, R. J., E. Granot, Y. Oschry, L. Rose and S. Eisenberg (1984). "Plasma triglyceride determines structure-composition in low and high density lipoproteins." Arteriosclerosis **4**(3): 225-231.

Endo, S. and K. U. Goss (2011). "Serum albumin binding of structurally diverse neutral organic compounds: data and models." Chem Res Toxicol **24**(12): 2293-2301.

Fischer, F. C., L. Henneberger, M. Konig, K. Bittermann, L. Linden, K. U. Goss and B. I. Escher (2017). "Modeling Exposure in the Tox21 *in vitro* Bioassays." Chem Res Toxicol **30**(5): 1197-1208.

Fisher, C., I. Gardner and J. Masoud (2017). "VIVD: A virtual *in vitro* distribution model for predicting intra- and sub-cellular concentrations in toxicity assays." Toxicology Letters **280**(Supplement 1):S290.

Fisher, C., M. Jamei and I. Gardner (2017). "VIVD: A virtual *in vitro* distribution model for predicting intra- and sub-cellular concentrations in toxicity assays." Toxicology Letters **280**: S290.

Gulden, M. and H. Seibert (2003). "*In vitro-in vivo* extrapolation: estimation of human serum concentrations of chemicals equivalent to cytotoxic concentrations *in vitro*." Toxicology **189**(3): 211-222.

Harwood, M. D., B. Achour, S. Neuhoff, M. R. Russell, G. Carlson, G. Warhurst and R. -H. Amin (2016). "*In vitro-In vivo* Extrapolation Scaling Factors for Intestinal P-Glycoprotein and Breast Cancer Resistance Protein: Part I: A Cross-Laboratory Comparison of Transporter-Protein Abundances and Relative Expression Factors in Human Intestine and Caco-2 Cells." Drug Metab Dispos **44**(3): 297-307.

Harwood, M. D., B. Achour, S. Neuhoff, M. R. Russell, G. Carlson, G. Warhurst and A. Rostami - Hodjegan (2016). "*In vitro-In vivo* Extrapolation Scaling Factors for Intestinal P-glycoprotein and Breast Cancer Resistance Protein: Part II. The Impact of Cross-Laboratory Variations of Intestinal Transporter Relative Expression Factors on Predicted Drug Disposition." Drug Metab Dispos **44**(3): 476-480.

Harwood, M. D., S. Neuhoff, G. L. Carlson, G. Warhurst and A. Rostami-Hodjegan (2013). "Absolute abundance and function of intestinal drug transporters: a prerequisite for fully mechanistic *in vitro-in vivo* extrapolation of oral drug absorption." Biopharm Drug Dispos **34**(1): 2-28.

Jamei, M., G. L. Dickinson and A. Rostami-Hodjegan (2009). "A framework for assessing interindividual variability in pharmacokinetics using virtual human populations and integrating general knowledge of physical chemistry, biology, anatomy, physiology and genetics: A tale of 'bottom-up' vs 'top-down' recognition of covariates." Drug Metab Pharmacokinet **24**(1): 53-75.

Jamei, M., S. Marciniak, D. Edwards, K. Wragg, K. Feng, A. Barnett and A. Rostami-Hodjegan (2013). "The simcyp population based simulator: architecture, implementation, and quality assurance." *In silico* Pharmacol **1**: 9.

Jamei, M., D. Turner, J. Yang, S. Neuhoff, S. Polak, A. Rostami-Hodjegan and G. Tucker (2009). "Population-based mechanistic prediction of oral drug absorption." AAPS J **11**(2): 225-237.

Kameya, H., N. Hokama, N. Hobara, S. Ohshiro and T. Uno (2009). "Effects of a dopamine receptor agonist and atropine sulfate on absorption of valproic acid in rats." Biomed Res **30**(2): 101-106.

Kazmi, F., T. Hensley, C. Pope, R. S. Funk, G. J. Loewen, D. B. Buckley and A. Parkinson (2013). "Lysosomal sequestration (trapping) of lipophilic amine (cationic amphiphilic) drugs in immortalized human hepatocytes (Fa2N-4 cells)." Drug Metab Dispos **41**(4): 897-905.

Kilford, P. J., M. Gertz, J. B. Houston and A. Galetin (2008). "Hepatocellular binding of drugs: correction for unbound fraction in hepatocyte incubations using microsomal binding or drug lipophilicity data." Drug Metab Dispos **36**(7): 1194-1197.

Kramer, N. I. (2010). Measuring, modelling and increasing the free concentration of test chemicals in cell assays. PhD, Utrecht University.

Kupke, D. W., M. G. Hodgins and J. W. Beams (1972). "Simultaneous determination of viscosity and density of protein solutions by magnetic suspension." Proc Natl Acad Sci U S A **69**(8): 2258-2262.

Loscher, W. (1978). "Serum protein binding and pharmacokinetics of valproate in man, dog, rat and mouse." J Pharmacol Exp Ther **204**(2): 255-261.

McGinnity, D. F., A. J. Berry, J. R. Kenny, K. Grime and R. J. Riley (2006). "Evaluation of timedependent cytochrome P450 inhibition using cultured human hepatocytes." Drug Metab Dispos **34**(8): 1291-1300.

Musther, H., M. D. Harwood, J. Yang, D. B. Turner, A. Rostami-Hodjegan and M. Jamei (2017). "The Constraints, Construction, and Verification of a Strain-Specific Physiologically Based Pharmacokinetic Rat Model." J Pharm Sci **106**(9): 2826-2838.

Nilsson, D., U. Fagerholm and H. Lennernas (1994). "The influence of net water absorption on the permeability of antipyrine and levodopa in the human jejunum." Pharm Res **11**(11): 1540-1547.

Poulin, P. and F. P. Theil (2002). "Prediction of pharmacokinetics prior to *in vivo* studies. 1. Mechanism-based prediction of volume of distribution." J Pharm Sci **91**(1): 129-156.

Rodgers, T., D. Leahy and M. Rowland (2005). "Physiologically based pharmacokinetic modeling 1: predicting the tissue distribution of moderate-to-strong bases." J Pharm Sci **94**(6): 1259-1276.

Rodgers, T. and M. Rowland (2006). "Physiologically based pharmacokinetic modelling 2: predicting the tissue distribution of acids, very weak bases, neutrals and zwitterions." J Pharm Sci **95**(6): 1238-1257.

Rodgers, T. and M. Rowland (2007). "Physiologically-based Pharmacokinetic Modeling 2: Predicting the tissue distribution of acids, very weak bases, neutrals and zwitterions." J Pharm Sci **95**: 1238-1257.

Sawada, Y., M. Hanano, Y. Sugiyama, H. Harashima and T. Iga (1984). "Prediction of the volumes of distribution of basic drugs in humans based on data from animals." J Pharmacokinet Biopharm **12**(6): 587-596.

Singh, K., J. M. Orr and F. S. Abbott (1988). "Pharmacokinetics and enterohepatic circulation of 2-npropyl-4-pentenoic acid in the rat." Drug Metab Dispos **16**(6): 848-852.

Sun, D., H. Lennernas, L. S. Welage, J. L. Barnett, C. P. Landowski, D. Foster, D. Fleisher, K. D. Lee and G. L. Amidon (2002). "Comparison of human duodenum and Caco-2 gene expression profiles for 12,000 gene sequences tags and correlation with permeability of 26 drugs." Pharm Res **19**(10): 1400-1416.

Trapp, S., G. R. Rosania, R. W. Horobin and J. Kornhuber (2008). "Quantitative modeling of selective lysosomal targeting for drug design." Eur Biophys J **37**(8): 1317-1328.

Wang, J. and D. R. Flanagan (1999). "General solution for diffusion-controlled dissolution of spherical particles. 1. Theory." J Pharm Sci **88**(7): 731-738.

Wang, J. and D. R. Flanagan (2002). "General solution for diffusion-controlled dissolution of spherical particles. 2. Evaluation of experimental data." J Pharm Sci **91**(2): 534-542.

Winiwarter, S., F. Ax, H. Lennernas, A. Hallberg, C. Pettersson and A. Karlen (2003). "Hydrogen bonding descriptors in the prediction of human *in vivo* intestinal permeability." J Mol Graph Model **21**(4): 273-287.

Winiwarter, S., N. M. Bonham, F. Ax, A. Hallberg, H. Lennernas and A. Karlen (1998). "Correlation of human jejunal permeability (*in vivo*) of drugs with experimentally and theoretically derived parameters. A multivariate data analysis approach." J Med Chem **41**(25): 4939-4949.

Yu, L. X. and G. L. Amidon (1998). "Saturable small intestinal drug absorption in humans: modelling and interpretation of cefatrizine data." Eur J Pharm Biopharm **45**(2): 199-203.

Zaldivar Comenges, J. M., E. Joossens, J. V. S. Benito, A. Worth and A. Paini (2017). "Theoretical and mathematical foundation of the Virtual Cell Based Assay - A review." Toxicol *In vitro* **45**(Pt 2): 209-221.

5. Plasma protein binding – description of 4 different models

Model 1: Random forest model for predicting the fraction unbound (square root of fraction unbound) to plasma proteins

QMRF identifier (JRC Inventory):To be entered by JRC

QMRF Title:Random forest model for predicting the fraction unbound (square root of fraction unbound) to plasma proteins

Printing Date:26-giu-2018

1. QSAR identifier

1.1. QSAR identifier (title):

Random forest model for predicting the fraction unbound (square root of fraction unbound) to plasma proteins

1.2. Other related models:

Random forest model for predicting the fraction unbound (square root of fraction unbound) of acidic drugs to plasma proteins

CORAL model for predicting the fraction unbound to plasma proteins (Square Root Transformation).

1.3. Software coding the model:

KNIME

Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization

Prof. Dr. Michael Berthold, Michael.Berthold@uni-konstanz.de https://www.knime.com/

2. General information

2.1. Date of QMRF:

25 June 2018

2.2. QMRF author(s) and contact details:

[1] Cosimo Toma; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri cosimo.toma@marionegri.it

[2] Domenico Gadaleta; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri domenico.gadaleta@marionegri.it

[2] Emilio Benfenati; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri emilio.benfenati@marionegri.it

2.3. Date of QMRF update(s):

2.4. QMRF update(s):

2.5. Model developer(s) and contact details:

Cosimo Toma IRCCS - Istituto di Ricerche Farmacologiche Mario Negri cosimo.toma@marionegri.it

2.6. Date of model development and/or publication:

2018

2.7. Reference(s) to main scientific papers and/or software package:

[1] Berthold, Michael R., *et al.* "KNIME-the Konstanz information miner: version 2.0 and beyond." AcM SIGKDD explorations Newsletter 11.1 (2009): 26-31. https://doi.org/10.1145/1656274.1656280

[2] Malot, C., VSURF: An R Package for Variable Selection Using Random Forests. The R Journal 2015, 7, 19-33.

[3] Chemaxon (2017). JChem for Office (Excel). JChem for Office. Collaborative Drug Discovery, I. (2010). ChemCell - Cheminformatics Workflow Automation for Microsoft Excel https://chemaxon.com/

2.8. Availability of information about the model:

All parts of the model freely available except for the calculation of ionisation state, that requires a ChemAxon academic license (https://chemaxon.com/)

2.9. Availability of another QMRF for exactly the same model:

3. Defining the endpoint - OECD Principle 1

3.1. Species: Human

3.2. Endpoint:

QMRF 5. Toxicokinetics QMRF 5. 9. Toxicokinetics.Protein-binding

3.3. Comment on endpoint:

Collection of protein plasma binding *in vivo* data (fraction unbound) from different literature sources.

3.4. Endpoint units:

Adimensional

3.5. Dependent variable:

For modelling purposes the endpoint was transformed in square root of fraction unbound (√fu).

3.6. Experimental protocol:

Described in Obach *et al.*, 2008, Drug Metab Dispos 36(7): 1385-1405

3.7. Endpoint data quality and variability:

4. Defining the algorithm - OECD Principle 2

4.1. Type of model:

Random Forest

4.2. Explicit algorithm:

Random Forest

Each tree was derived from a random sampling with replacement of training set data. The attributes for each tree were randomly selected from the initial pool of descriptors. The number of attributes of each tree was the square root of the initial number of descriptors.

The number of trees is 100.

Ionisation state of compounds was defined before descriptors calculation and model derivation. Ionisation state was calculated with JChem extension for Microsoft Excel provided by ChemAxon.

4.3. Descriptors in the model:

[1] ALOGP Ghose-Crippen octanol-water partition coeff. (logP)

[2] C% percentage of C atoms

[3] C-024 R--CH--R

[4] CATS2D_00_LL CATS2D Lipophilic-Lipophilic at lag 00

[5] CATS2D_00_PP CATS2D Positive-Positive at lag 00

[6] CATS2D_01_LL CATS2D Lipophilic-Lipophilic at lag 01

[7] Eta_betaP eta pi and lone pair VEM count

[8] Eta_betaP_A eta pi and lone pair average VEM count

[9] MLOGP Moriguchi octanol-water partition coeff. (logP)

[10] N% percentage of N atoms

[11] PCD difference between multiple path count and path count

[12] P_VSA_i_2 P_VSA-like on ionisation potential, bin 2

[13] P_VSA_p_3 P_VSA-like on polarizability, bin 3

[14] SM12_AEA(ri) spectral moment of order 12 from augmented edge adjacency mat. Weighted by resonance integral

[15] SpMax2_Bh(m) largest eigenvalue n. 2 of Burden matrix weighted by mass

[16] SpMin1_Bh(i) smallest eigenvalue n. 1 of Burden matrix weighted by ionisation potential

[17] Ui unsaturation index

[18] nCar number of aromatic C(sp2)

[19] nN+ number of positively charged N

[20] totalcharge total charge

4.4. Descriptor selection:

The initial pool included 3850 2D descriptors calculated with Dragon 7.0. Descriptors were filtered based on 1) Variance, 2) Absolute Pair Correlation and 3) VSURF (R package).

4.5. Algorithm and descriptor generation:

Feature selection was based on training set chemicals. Descriptors were pruned by constant and semi-constant vales (i.e. standard deviation < 0.01), then if a couple of descriptors was characterised by an absolute pair correlation greater than 90%, the descriptor with the highest pair correlation with all the other descriptors was removed. Optimal subsets of descriptors for modelling were obtained with the R package VSURF. The algorithm consists in a three step variable selection based on the logic underpinning the random forest (RF) algorithm (i.e. permutation importance and out-of-bag error). The first step eliminates irrelevant descriptors according to the permutation-based RF score of importance and a user-defined threshold. The second step finds important descriptors closely related to the response variable (interpretation step) and the third step (prediction step) identifies a sufficient parsimonious set of important descriptors leading to a good prediction of the response variables.

4.6. Software name and version for descriptor generation:

Dragon 7.0

Calculation of several sets of molecular descriptors from molecular geometries (topological, geometrical, WHIM, 3D-MoRSE, molecular profiles, etc.)

Kode srl. Via Nino Pisano, 14 56122 Pisa (PI) - Italy, info@kode-solutions.net www.kodesolutions.net

https://chm.kode-solutions.net/products_dragon.php

4.7. Chemicals/Descriptors ratio:

391/20

5. Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model:

Two-class real-random classification

5.2. Method used to assess the applicability domain:

Descriptors of training set chemicals are randomly permuted (vert ical permutation) to create a mirror training set. The shuffled training set are merged with the original one. Samples or the original training set are flagged as “real”, while those of the mirror training set are flagged as “dummy”. A classification model (Random Forest) is built to distinguish

real from dummy samples. External chemicals classified as “dummy” are considered outside of the model’s Applicability Domain.

5.3. Software name and version for applicability domain assessment:

KNIME

Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization

Prof. Dr. Michael Berthold, Michael.Berthold@uni-konstanz.de https://www.knime.com/

5.4. Limits of applicability:

The model is not applicable on inorganic chemicals and those including unusual elements (i.e., different from C, O, N, S, Cl, Br, F, I). Salts can be predicted only if stripped of the counterion and converted to the neutralised form.

Compounds for which ALOGP cannot be calculated are considered outside of the applicability domain.

6. Internal validation - OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Available information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable for the training set:

All

6.5. Other information about the training set:

6.6. Pre-processing of data before modelling:

SMILES were retrieved using different chemicals identifiers (chemical name, CAS number) checking data from different web sources (i.e., EPA’s CompTox database, ChemIDPlus, PubChem) by mean of an automated in-house tool. SMILES were stripped of their counterions and neutralised.

Chemicals were checked for removal of inorganic chemicals and mixtures, and for correction of inaccurate SMILES codes with the help of chemical databases.

Descriptors were centered and autoscaled before modelling.

Fraction unbound data were converted to square root of fraction unbound (√fu) for modelling purposes.

6.7. Statistics for goodness-of-fit:

6.8. Robustness - Statistics obtained by leave-one-out cross-validation:

6.9. Robustness - Statistics obtained by leave-many-out cross-validation:

Method: 10-fold cross-validation:

$R^2 = 0.60$

RMSE = 0.19

6.10. Robustness - Statistics obtained by Y-scrambling:

6.11. Robustness - Statistics obtained by bootstrap:

6.12. Robustness - Statistics obtained by other methods:

7. External validation - OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Available information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

All

7.4. Data for the dependent variable for the external validation set:

All

7.5. Other information about the external validation set:

7.6. Experimental design of test set:

Activity sampling method. The entire dataset was sorted based on activity and divided in equal sized bins. For each bin, the 80% of chemicals were assigned to the training set while the 20% was assigned to the validation set.

7.7. Predictivity - Statistics obtained by external validation:

r^2= 0.72

RMSE= 0.16

Percentage of chemicals within the applicability domain = 87%

7.8. Predictivity - Assessment of the external validation set:

The activity sampling method allowed to design a validation set that chemically representative of training set compounds.

7.9. Comments on the external validation of the model:

8. Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model:

Plasma protein binding is heavily influenced by lipophilicity and ionisation of compounds.

8.2. A priori or a posteriori mechanistic interpretation:

8.3. Other information about the mechanistic interpretation:

9. Miscellaneous information

9.1. Comments:

9.2. Bibliography:

[1] Obach, R. S., F. Lombardo and N. J. Waters (2008). "Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds." Drug Metab Dispos 36(7): 1385-1405.

[2] Hastie, T. (2008). Tibshirani, R. and Friedman, J.(2009): The elements of statistical learning. Data mining, inference, and prediction, Springer, New York, ISBN.

[3] Kuhn, M. and K. Johnson (2013). Applied Predictive Modelling, Springer-Verlag New York: XIII, 600.

9.3. Supporting information:

Training set(s)Test set(s)Supporting information

Model 2: CORAL model for predicting fraction unbound to plasma proteins (Square root Transformation)

QMRF identifier (JRC Inventory):To be entered by JRC

QMRF Title: Coral model for predicting of fraction unbound to plasma proteins (Square root Transformation)

Printing Date:26-giu-2018

1. QSAR identifier

1.1. QSAR identifier (title):

CORAL model for predicting fraction unbound to plasma proteins (Square root Transformation)

1.2. Other related models:

Random Forest model for predicting of fraction unbound of plasma protein binding (Square root Transformation)

Random forest model for predicting the fraction unbound (square root of fraction unbound) of acidic drugs to plasma proteins

1.3. Software coding the model: Coral(CORrelation And Logic)

CORAL breaks the chemical structures of the compounds in the training set into small components (SMILES attributes), based on the SMILES structure in the canonical form.

Andrey Toropov, andrey.toropov@marionegri.it; Emilio Benfenati emilio.benfenati@marionegri.it http://www.insilico.eu/coral/SOFTWARECORAL.html

2. General information

2.1. Date of QMRF:

25 June 2018

2.2. QMRF author(s) and contact details:

[1] Cosimo Toma IRCCS - Istituto di Ricerche Farmacologiche Mario Negri cosimo.toma@marionegri.it

[2] Emilio Benfenati IRCCS - Istituto di Ricerche Farmacologiche Mario Negri emilio.benfenati@marionegri.it

[3] Andrey Toropov IRCCS - Istituto di Ricerche Farmacologiche Mario Negri andrey.toropov@marionegri.it

[4] Domenico Gadaleta; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri domenico.gadaleta@marionegri.it

2.3. Date of QMRF update(s):

2.4. QMRF update(s):

2.5. Model developer(s) and contact details:

Andrey Toropov IRCCS - Istituto di Ricerche Farmacologiche Mario Negri andrey.toropov@marionegri.it

2.6. Date of model development and/or publication:

2018

2.7. Reference(s) to main scientific papers and/or software package:

A.P. Toropova, A.A. Toropova, R. Gonella Diaza, E. Benfenati, G. Gini, Analysis of the co-evolutions of correlations as a tool for QSAR-modelling carcinogenicity: an unexpected good prediction based on a model that seems untrustworthy. //Cent. Eur. J. Chem. 9 (2011) 165-174

2.8. Availability of information about the model:

All parts of the model freely avaliable except for the ionisation state calculation, that requires ChemAxon academic licence (https://chemaxon.com/)

2.9. Availability of another QMRF for exactly the same model:

3. Defining the endpoint - OECD Principle 1

3.1. Species: Human

3.2. Endpoint:

QMRF 5. Toxicokinetics QMRF 5. 9. Toxicokinetics.Protein-binding

3.3. Comment on endpoint:

Collection of *in vivo* data from literature of plasma protein binding (fraction unbound data)

3.4. Endpoint units:

Adimensional

3.5. Dependent variable:

The endpoint is transformed in square root of fraction unbound (√fu) for modelling purposes

3.6. Experimental protocol:

Described in Obach *et al.*, 2008

3.7. Endpoint data quality and variability:

4. Defining the algorithm - OECD Principle 2

4.1. Type of model:

Least Square

4.2. Explicit algorithm:

In CORAL models, the endpoint is a function of these SMILES-based descriptors:

$$\text{Endpoint} = C0 + C1\ \text{DCW}(T, N)$$

where C0 and C1 are the intercept and slope for the relationship, and DCW (T, N) is the combination of SMILES-based attributes, each associated with a correlation weight (CW). CWs are determined with the Monte Carlo algorithm in an iterative procedure that aims to optimise a target function (TF):

$$TF = R + R' - |R - R'|\ 0.01$$

where R and R' are the correlation coefficients between DCW(T, N) and the endpoints for the training set and internal training set. This procedure is defined as a balance of correlations (BC). The TF is a function of the CWs and is optimised by iteratively modifying them. In the first part of the optimisation, CWs are incremented by a value Dstart. This increment is repeated as long as these was a corresponding improvement of the TF. When no further improvement is observed, the Dstart value is modified to Dstart,1 = -0.5 (Dstart) for subsequent iterations. Dstar is iteratively modified each time that an increment of CWs fails to correspond to an increment of TF, until | Dstart | is lower than a threshold value (Dprecession). N is the number of epochs of Monte Carlo for optimisation of the TF, and T is a threshold used to classify SMILES attributes as rare or not rare. An attribute is defined as rare if it is found in the SMILES of the calibration set less than T times. Rare SMILES attribute values were set to zero so they were not involved in the modelling. T and N are set to optimise the statistical performance for the calibration set.

T = 1; N = 5; D_{start}= 1.5; $D_{precession}$ = 0.1

4.3. Descriptors in the model:

[1] SSk Combination of two SMILES elements (family of descriptors)

[2] SSSk Combination of three SMILES elements (family of descriptors)

The optimal descriptors calculated with the CORAL software are attributes extracted from molecules represented as SMILES notation, which check the presence of particular characters (or combinations of characters) within the SMILES.

Sk, SSk are SMILES attributes defined by a sequence of atoms and bonds present in the SMILES string. Sk represents single elements, and SSk two elements combined. Attributes with a positive CW are considered promoters of an increase of the endpoint value, while attributes with a negative correlation weights are considered promoters of a decrease.

4.4. Descriptor selection:

Monte Carlo optimisation (see 4.2).

4.5. Algorithm and descriptor generation:

CORAL breaks the chemical structures of the compounds in the training set into small components (SMILES attributes), based on the SMILES structure in the canonical form.

4.6. Software name and version for descriptor generation:

Coral(CORrelation And Logic)

CORAL breaks the chemical structures of the compounds in the training set into small components (SMILES attributes), based on the SMILES structure in the canonical form.

Andrey Toropov, andrey.toropov@marionegri.it; Emilio Benfenati emilio.benfenati@marionegri.it http://www.insilico.eu/coral/SOFTWARECORAL.html

4.7. Chemicals/Descriptors ratio:

5. Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model:

CORAL in-house applicability domain

5.2. Method used to assess the applicability domain:

Only compounds whose SMILES present attributes included among those that constitute the model are inside the AD.

5.3. Software name and version for applicability domain assessment:

Coral (CORrelation And Logic)

CORAL breaks the chemical structures of the compounds in the training set into small components (SMILES attributes), based on the SMILES structure in the canonical form.

Andrey Toropov, andrey.toropov@marionegri.it; Emilio Benfenati emilio.benfenati@marionegri.it http://www.insilico.eu/coral/SOFTWARECORAL.html

5.4. Limits of applicability:

The model is not applicable to inorganic chemicals and those including unusual elements (i.e., different from C, O, N, S, Cl, Br, F, I). Salts can be predicted only if stripped of the counterion and converted to the neutralised form.

6. Internal validation - OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Available information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable for the training set:

All

6.5. Other information about the training set:

The initial dataset is randomly divided into a Training set (TS), an Invisible Training set (ITS), a Calibration set (CS) and a Validation set (VS). TS, ITS and CS were used for model training and calibration, and VS for external validation.

6.6. Pre-processing of data before modelling:

SMILES were retrieved using different chemicals identifiers (chemical name, CAS number) checking data from different web sources (i.e., EPA's CompTox database, ChemIDPlus, PubChem) by mean of an automated in-house tool. SMILES were stripped of their counterions and neutralised.

Chemicals were checked for removal of inorganic chemicals and mixtures, and for correction of inaccurate SMILES codes with the help of chemical databases.

Fraction unbound data were converted to square root of fraction unbound (√fu) for modelling purposes.

6.7. Statistics for goodness-of-fit:

Performance on training set:

r^2 =0.61 RMSE = 0.19

6.8. Robustness - Statistics obtained by leave-one-out cross-validation:

6.9. Robustness - Statistics obtained by leave-many-out cross-validation:

6.10. Robustness - Statistics obtained by Y-scrambling:

6.11. Robustness - Statistics obtained by bootstrap:

6.12. Robustness - Statistics obtained by other methods:

7. External validation - OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Available information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

All

7.4. Data for the dependent variable for the external validation set:

All

7.5. Other information about the external validation set:

7.6. Experimental design of test set:

Random selection (see 6.5)

7.7. Predictivity - Statistics obtained by external validation:

Performance on the Validation set:

r^2= 0.65

RMSE= 0.18

Percentage of compounds within the AD = 91%

7.8. Predictivity - Assessment of the external validation set:

7.9. Comments on the external validation of the model:

8. Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model:

8.2. A priori or a posteriori mechanistic interpretation:

8.3. Other information about the mechanistic interpretation:

9. Miscellaneous information

9.1. Comments:

9.2. Bibliography:

Obach, R. S., F. Lombardo and N. J. Waters (2008). "Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds." Drug Metab Dispos 36(7): 13851405.

9.3. Supporting information:

Training set(s)Test set(s)Supporting information

Model 3: Random forest model for predicting the fraction unbound (square root of fraction unbound) of acidic drugs to plasma proteins

QMRF identifier (JRC Inventory):To be entered by JRC

QMRF Title: Random forest model for predicting the fraction unbound (square root of fraction unbound) of acidic drugs to plasma proteins

Printing Date:27-giu-2018

1. QSAR identifier

1.1. QSAR identifier (title):

Random forest model for predicting the fraction unbound (square root of fraction unbound) of acidic drugs to plasma proteins

1.2. Other related models:

Random forest model for predicting the fraction unbound (square root of fraction unbound) to plasma proteins

CORAL model for predicting the fraction unbound to plasma proteins (Square Root Transformation).

1.3. Software coding the model:

KNIME

Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization

Prof. Dr. Michael Berthold, Michael.Berthold@uni-konstanz.de https://www.knime.com/

2. General information

2.1. Date of QMRF:

25 June 2018

2.2. QMRF author(s) and contact details:

[1]Cosimo Toma IRCCS - Istituto di Ricerche Farmacologiche Mario Negri cosimo.toma@marionegri.it

[2] Domenico Gadaleta; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri domenico.gadaleta@marionegri.it

[2]Emilio Benfenati IRCCS - Istituto di Ricerche Farmacologiche Mario Negri emilio.benfenati@marionegri.it

2.3. Date of QMRF update(s):

2.4. QMRF update(s):

2.5. Model developer(s) and contact details:

Cosimo Toma IRCCS - Istituto di Ricerche Farmacologiche Mario Negri cosimo.toma@marionegri.it

2.6. Date of model development and/or publication:

2018

2.7. Reference(s) to main scientific papers and/or software package:

[1] Berthold, Michael R., *et al.* "KNIME-the Konstanz information miner: version 2.0 and beyond." AcM SIGKDD explorations Newsletter 11.1 (2009): 26-31. https://doi.org/10.1145/1656274.1656280

[2] Advanced Chemistry Development, I. (2010). ACD/labs. Toronto, ON, Canada.

2.8. Availability of information about the model:

All parts of the model freely available

2.9. Availability of another QMRF for exactly the same model:

3. Defining the endpoint - OECD Principle 1

3.1. Species: Human

3.2. Endpoint:

QMRF 5. Toxicokinetics QMRF 5. 9. Toxicokinetics.Protein-binding

3.3. Comment on endpoint:

Collection of protein plasma binding *in vivo* data (fraction unbound) from different literature sources.

3.4. Endpoint units:

Adimensional

3.5. Dependent variable:

For modelling purposes the endpoint was transformed in square root of fraction unbound (√fu).

3.6. Experimental protocol:

Described in Obach *et al.*, 2008, Drug Metab Dispos 36(7): 1385-1405

3.7.Endpoint data quality and variability:

4.Defining the algorithm - OECD Principle 2

4.1. Type of model:

Random Forest

4.2. Explicit algorithm:

Random Forest

Each tree was derived from a random sampling with replacement of training set data. The attributes for each tree were randomly selected from the initial pool of descriptors. The number of attributes of each tree was the square root of the initial number of descriptors.

The number of trees is 100.

4.3. Descriptors in the model:

[1]ALOGP Ghose-Crippen octanol-water partition coeff. (logP)

[2]ALOGP2 squared Ghose-Crippen octanol-water partition coeff. (logP^2)

[3]PCR ratio of multiple path count over path count

[4]MATS1e Moran autocorrelation of lag 1 weighted by Sanderson electronegativity

[5]SpMAD_EA(dm) spectral mean absolute deviation from edge adjacency mat. weighted by dipole moment

[6]nCar number of aromatic C(sp2)

[7]P_VSA_ppp_ar P_VSA-like on potential pharmacophore points, ar - aromatic atoms

[8]JGI3 mean topological charge index of order 3

4.4. Descriptor selection:

The initial pool included 3850 2D descriptors calculated with Dragon 7.0. Descriptors were filtered based on 1) Variance, 2) Absolute Pair Correlation and 3) VSURF (R package).

4.5. Algorithm and descriptor generation:

Feature selection was based on training set chemicals. Descriptors were pruned by constant and semi-constant vales (i.e. standard deviation < 0.01), then if a couple of descriptors was characterised by an absolute pair correlation greater than 90%, the descriptor with the highest pair correlation with all the other descriptors was removed. Optimal subsets of descriptors for modelling were obtained with the R package VSURF. The algorithm consists in a three step variable selection based on the logic underpinning the random forest (RF) algorithm (i.e. permutation importance and out -of-bag error). The first step eliminates irrelevant descriptors according to the permutation-based RF score of importance and a user-defined threshold. The second step finds important descriptors closely related to the response variable (interpretation step) and the third step (prediction step) identifies a sufficient parsimonious set of important descriptors leading to a good prediction of the response variables.

4.6.Software name and version for descriptor generation:

Dragon 7.0

Calculation of several sets of molecular descriptors from molecular geometries (topological, geometrical, WHIM, 3D-MoRSE, molecular profiles, etc.)

Kode srl. Via Nino Pisano, 14 56122 Pisa (PI) - Italy, info@kode-solutions.net www.kodesolutions.net

https://chm.kode-solutions.net/products_dragon.php

4.7. Chemicals/Descriptors ratio:

97 (training set)/8

5. Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model:

Two-class real-random classification

5.2. Method used to assess the applicability domain:

Descriptors of training set chemicals are randomly permuted (vertical permutation) to create a mirror training set. The shuffled training set are merged with the original one. Samples or the original training set are flagged as "real", while those of the mirror training set are flagged as "dummy". A classification model (Random Forest) is built to distinguish real from dummy samples. External chemicals classified as "dummy" are considered outside of the model's Applicability Domain.

5.3. Software name and version for applicability domain assessment:

KNIME

Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization

Prof. Dr. Michael Berthold, Michael.Berthold@uni-konstanz.de https://www.knime.com/

5.4. Limits of applicability:

The model is not applicable on inorganic chemicals and those including unusual elements (i.e., different from C, O, N, S, Cl, Br, F, I). Salts can be predicted only if stripped of the counterion and converted to the neutralised form.

The model is only applicable to organic acids. Salts should be stripped of the counterion and converted to the neutralised form before predictions.

Compounds for which ALOGP cannot be calculated are considered outside of the applicability domain.

6. Internal validation - OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Available information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

All

6.4. Data for the dependent variable for the training set:

All

6.5. Other information about the training set:

Only acidic compounds were considered for modelling, i.e. compounds being for more than 10% in the deprotonated (acidic) state concentration at pH 7.4.ACD/labs 12.0 (Advanced Chemistry Development 2010).

6.6. Pre-processing of data before modelling:

SMILES were retrieved using different chemicals identifiers (chemical name, CAS number) checking data from different web sources (i.e., EPA's CompTox database, ChemIDPlus, PubChem) by mean of an automated in-house tool. SMILES were stripped of their counterions and neutralised.

Chemicals were checked for removal of inorganic chemicals and mixtures, and for correction of inaccurate SMILES codes with the help of chemical databases.

Descriptors were centered and autoscaled before modelling.

Fraction unbound data were converted to square root of fraction unbound (√fu) for modelling purposes.

6.7. Statistics for goodness-of-fit:

6.8. Robustness - Statistics obtained by leave-one-out cross-validation:

6.9. Robustness - Statistics obtained by leave-many-out cross-validation:

Method: 10-fold cross-validation

$R^2 = 0.63$

RMSE = 0.20

6.10. Robustness - Statistics obtained by Y-scrambling:

6.11. Robustness - Statistics obtained by bootstrap:

6.12. Robustness - Statistics obtained by other methods:

7. External validation - OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Available information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

All

7.4. Data for the dependent variable for the external validation set:

All

7.5. Other information about the external validation set:

7.6. Experimental design of test set:

Activity sampling method. The entire dataset was sorted based on activity and divided in equal sized bins. For each bin, the 80% of chemicals were assigned to the training set while the 20% was assigned to the validation set.

7.7. Predictivity - Statistics obtained by external validation:

R^2= 0.73

RMSE= 0.17

Percentage of chemicals within the applicability domain = 96 %

7.8. Predictivity - Assessment of the external validation set:

The activity sampling method allowed to design a validation set that chemically representative of training set compounds.

7.9.Comments on the external validation of the model:

8. Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model:

Plasma protein binding is heavily influenced by lipophilicity and ionisation of compounds.

Acidic substances shows a high affinity for Albumin.

8.2. A priori or a posteriori mechanistic interpretation:

8.3. Other information about the mechanistic interpretation:

9. Miscellaneous information

9.1. Comments:

9.2. Bibliography:

[1] Obach, R. S., F. Lombardo and N. J. Waters (2008). "Trend analysis of a database of intravenous pharmacokinetic parameters in humans for 670 drug compounds." Drug Metab Dispos 36(7): 1385-1405.

[2] Hastie, T. (2008). Tibshirani, R. and Friedman, J.(2009): The elements of statistical learning. Data mining, inference, and prediction, Springer, New York, ISBN.

[3] Kuhn, M. and K. Johnson (2013). Applied Predictive Modelling, Springer-Verlag New York: XIII, 600.

9.3. Supporting information:

Training set(s)Test set(s)Supporting information

10. Summary (JRC QSAR Model Database)

10.1. QMRF number:

To be entered by JRC

10.2. Publication date:

To be entered by JRC

10.3. Keywords:

To be entered by JRC

10.4. Comments:

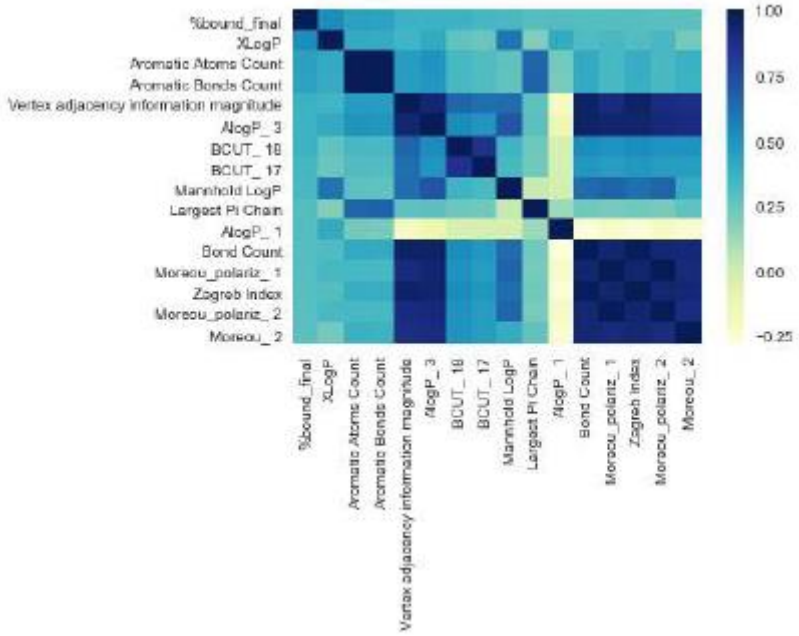
To be entered by JRC

Model 4: Lhasa model for ppb

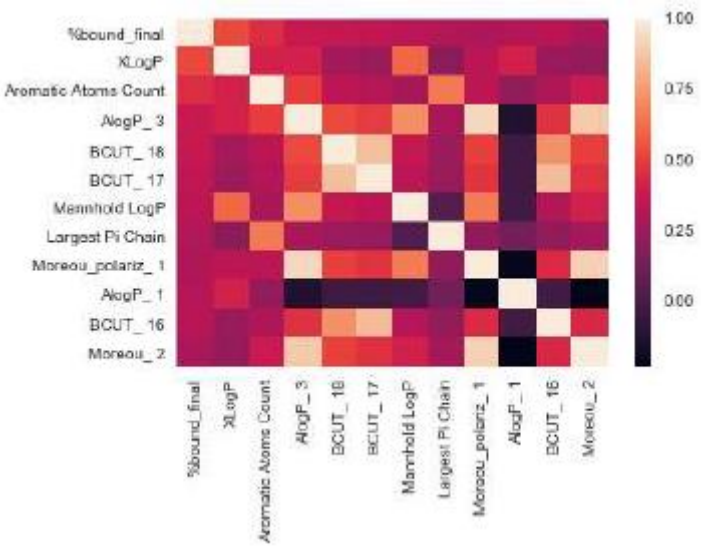
A plasma protein binding model based on the random forest algorithm was developed to predict the associated dissociation constant (Kd) calculated from log10 (Kd) = log10 (fb/fub). The predicted values for the case study are presented in Table 1 expressed in % bound obtained from Kd. The model was developed by 10 times cross-validation using a dataset of PPB collected from the literature and the ChEMBL dataset (3600 chemicals in total, where 20% was used as test sets). Using an in-house pKa model all chemicals were converted to their ionisation state at the pH= 7.4. The highly correlated descriptors were removed using a threshold of 0.95 (Figure 1).

Figure 1. Correlation analysis of the descriptors.

a) Correlation analysis before removal of the highly correlated features.



b) Correlation analysis after removal of the highly correlated features.



The features used for this model are from the CDK descriptor set. A dictionary of CDK descriptors is provided at: http://qsar.sourceforge.net/dicts/qsar-descriptors/index.xhtml. For example, definitions of the most relevant descriptors: BCUT, Moreau and Number of Rotatable bonds can be found in the references as follows:

BCUT definition: Eigenvalue based descriptor noted for its utility in chemical diversity described by Pearlman et al. [PEA99].

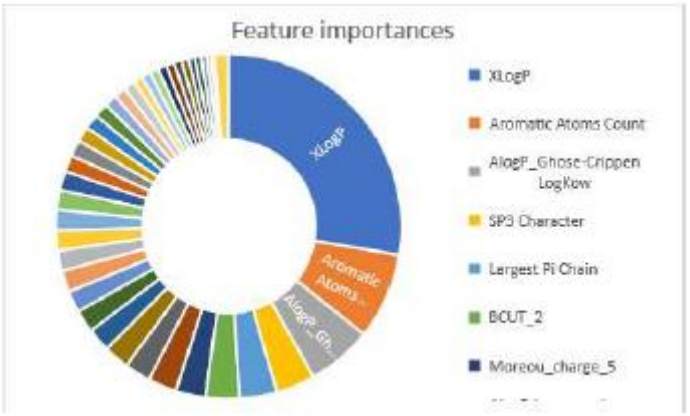
Moreau Broto Autocorrelation descriptor definition:
http://www.rguha.net/writing/notes/desc/desc.html

Number of rotatable bonds descriptors definition:
http://www.talete.mi.it/help/dproperties_help/index.html?constitutional_descriptors.htm

mechanistically it can explain the plasma protein binding of the chemical.

The random forest model feature importance rank is shown in Figure 2.

Figure 2. Feature importances based on the RF model



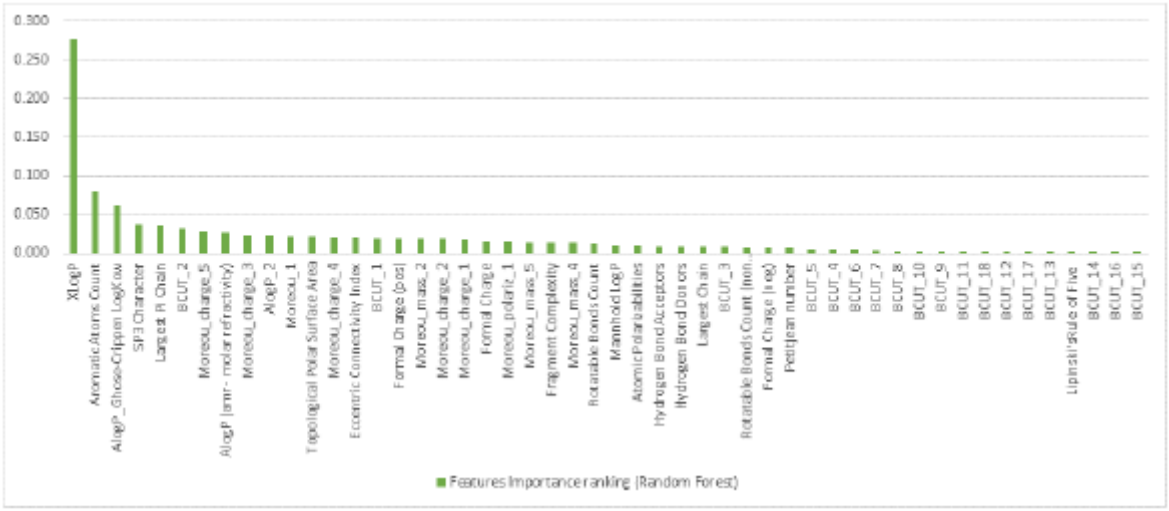
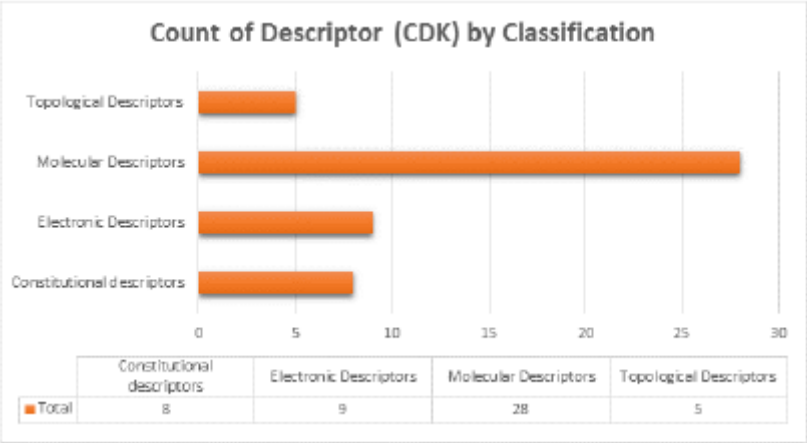


Figure 3 shows the count of descriptors used in the model by classification into molecular, topological, constitutional, and electronic descriptors.

Figure 3: count of descriptors by classification



Predictions of log10(Kd) performance on the test data were as follows: R^2 score = 0.61, MAE = 0.438, RMSE = 0.588. An applicability domain (AD) approach was developed which considers three steps: 1) chemical type (i.e. salts and inorganics are outside AD), 2) minimum and maximum values of the descriptors for the test chemical should to fall in the range of those of the chemicals in the training set, 3) a nearest neighbours approach comparing the test chemical descriptors to the cumulative values of the similarity in within the training set chemicals. Results of these analyses are presented in table 1.

Table 1. Results of the prediction of the %PPB and the AD assessment.

Name	CAS	Predicted_%PPB	Applicability domain	Reliability (distance)	Ionisation state at pH=7.4	Observed %bound
2-Ethylbutyric Acid	88-09-5	28.77	inside	0.004227	acid	
2-Propylheptanoic Acid	31080-39-4	73.73	inside	0.019021	acid	
2-Ethylheptanoic Acid	3274-29-1	76.63	inside	0.015851	acid	
2-Propylhexanoic Acid	3274-28-0	80.43	inside	0.015851	acid	
Valproic Acid	99-66-1	86.23	inside	0.179641	acid	91.6
2-Ethylhexanoic Acid	149-57-5	79.86	inside	0.123987	acid	
2-Ethylpentanoic Acid	20225-24-5	51.18	inside	0.015851	acid	
2-Methylbutyric Acid	1730-91-2	23.70	inside	0.002466	acid	
2-Methylpentanoic Acid	97-61-0	32.8698513	inside AD	0.004579	acid	
2-Methylhexanoic Acid	4536-23-6	54.55610557	inside AD	0.013033	acid	
Pivalic Acid	75-98-9	31.89740927	inside AD	0.002818	acid	

6. Metabolism Information for Mock Submission

Known Metabolism of Case Study Target and Analogues

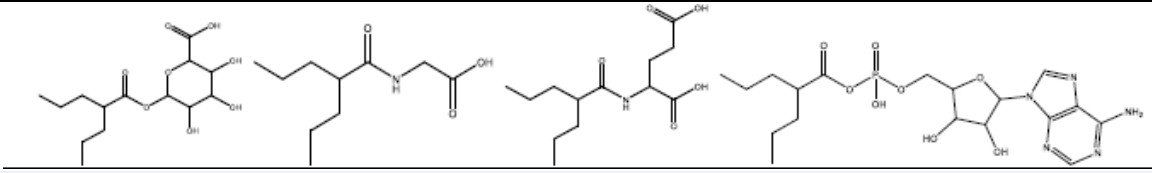
A search of the literature has unearthed variable amounts of information on the metabolism of four out of the ten chosen analogue compounds. Sources include **PubMed**, the **Human Metabolome Database, PubChem**, **DrugBank** as well as nonspecialised sources such as **Google**. Two of the ten analogue compounds (valproic acid and pivalic acid) have some representation in the **Lhasa Limited metabolism data set**. Analogue compounds with some known metabolism are: **valproic acid**, **2-ethylhexanoic acid**, **2-methylhexanoic acid** and **pivalic acid**.

Valproic acid

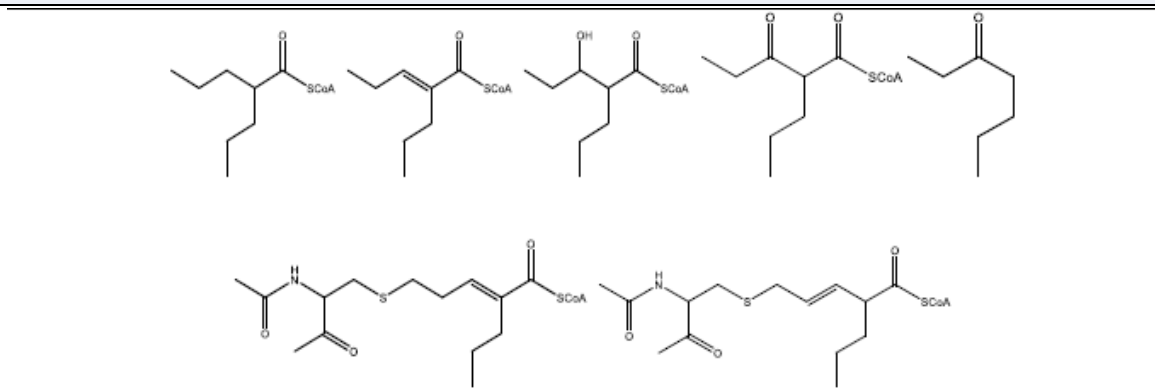
The *in vivo* and *in vitro* metabolism of valproic acid (VPA) and 4-ene VPA is extensive, well documented and reviewed[1, 2] for a number of animal species and human. Metabolites can be categorised as those arising from CYP-catalysed oxidation at alkyl side chains (microsomes, endoplasmic reticulum), those arising from beta-oxidation pathways (mitochondria) and those occurring as a consequence of phase II reactions of conjugation (endoplasmic reticulum and cytosol). Variability of VPA metabolism and pharmacokinetics in different disease states and co-administrations are well-studied. VPA is highly protein bound (87–95%) resulting in low clearance (6–20 ml/h/kg). Its major urinary metabolite is the valproate glucuronide accounting for 30-50% of an administered dose. beta-Oxidation is the most important oxidative biotransformation type (>40%) for VPA with CYP-based hydroxylation/dehydrogenation (15-20%) playing a secondary role. Some 50-70 different metabolites have been suggested in the literature for VPA. Some example metabolites in all three biotransformation categories are shown below.

Valproic Acid Metabolites

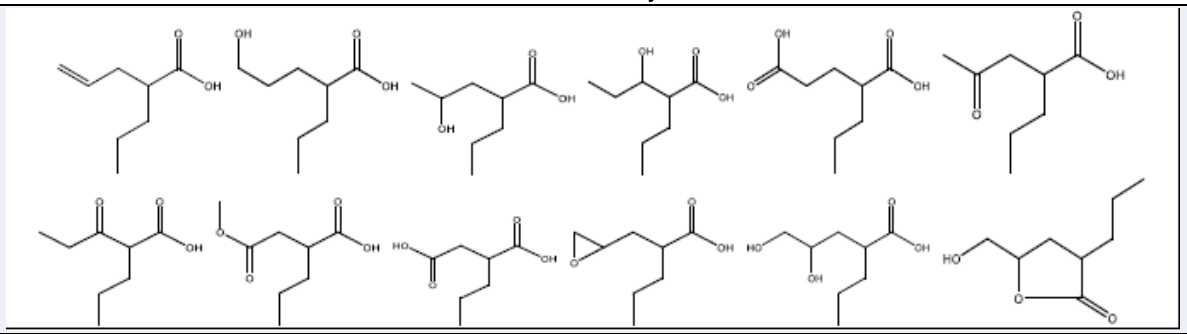
Metabolites from Conjugation Reactions



Metabolites from beta-Oxidation



Metabolites from CYP-Catalysed Oxidation

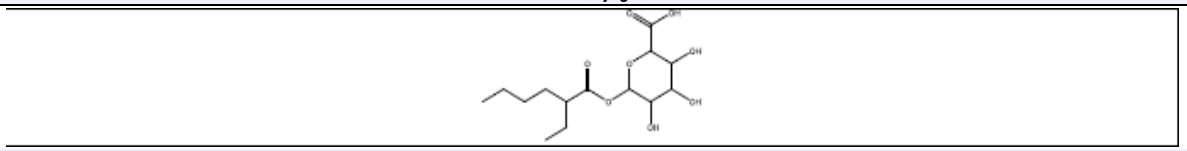


2-Ethylhexanoic acid

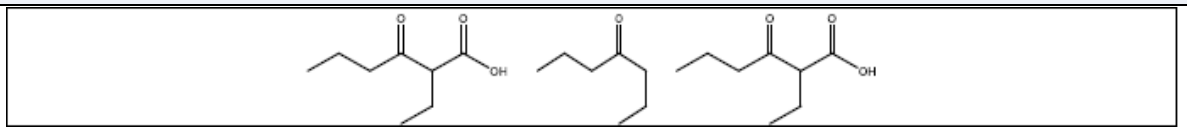
Along with 2-ethylhexanol, 2-ethylhexanoic acid is a metabolite of the plasticiser bis(2-ethylhexyl)phthalate and is reasonably well studied[3,4,5,6]. Data is available from both *in vitro* and *in vivo* experiments in human, rat, rabbit, mouse, monkey, guinea pig and dog. Like valproic acid (of which 2-ethylhexanoic acid is a chain isomer) metabolism is by glucuronide conjugation, beta-oxidation and CYP-mediated hydroxylation/dehydrogenation. Some example metabolites in all three biotransformation categories are shown below. Additional but uncharacterised hydroxylated metabolites and two lactones have also been reported.

2-Ethylhexanoic Acid Metabolites

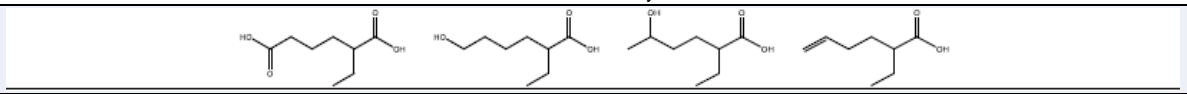
Metabolites from Conjugation Reactions



Metabolites from beta-Oxidation



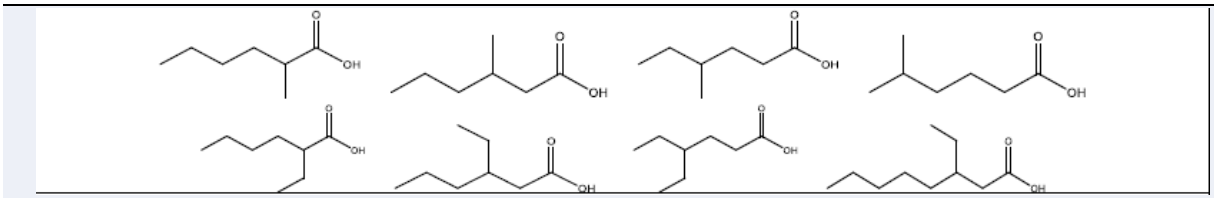
Metabolites from CYP-Catalysed Oxidation



2-Methylhexanoic acid

The metabolism of 2-methylhexanoic acid has been reported to be by glucuronidation in hepatic microsomes of several species including human[7]. Glucuronidation activity toward several analogue compounds (3-methylhexanoic acid, 4-methylhexanoic acid, 5-methylhexanoic acid, 2-ethylhexanoic acid, 3-ethylhexanoic acid, 4-ethylhexanoic acid and 3-ethyloctanoic acid) increased as a function of molecular weight, but was not affected by the position of the methyl or the ethyl moiety on the hydrocarbon chain.

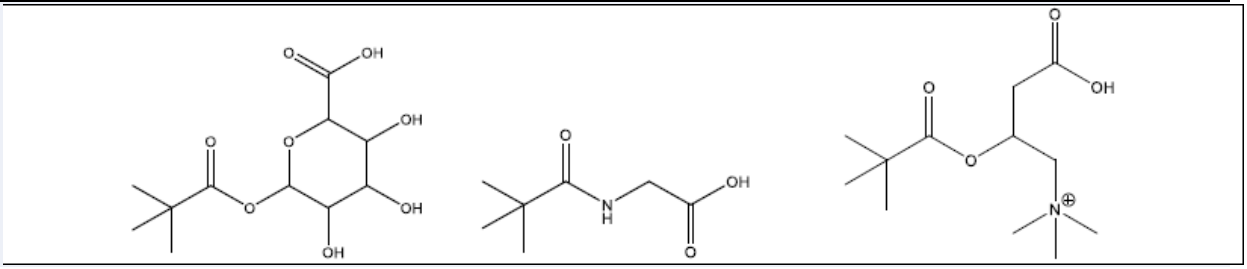
2-Methylhexanoic Acid and Analogues that Undergo Glucuronidation (from reference 7)



Pivalic acid

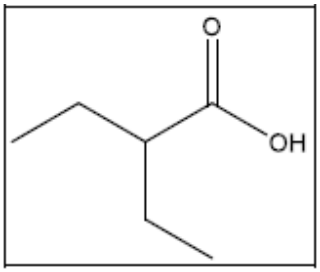
Pivalic acid undergoes conjugation with glucuronic acid, glycine and carnitine in rat, dog, rabbit and monkey hepatocytes and kidney slices[8, 9].

Pivalic Acid Metabolites

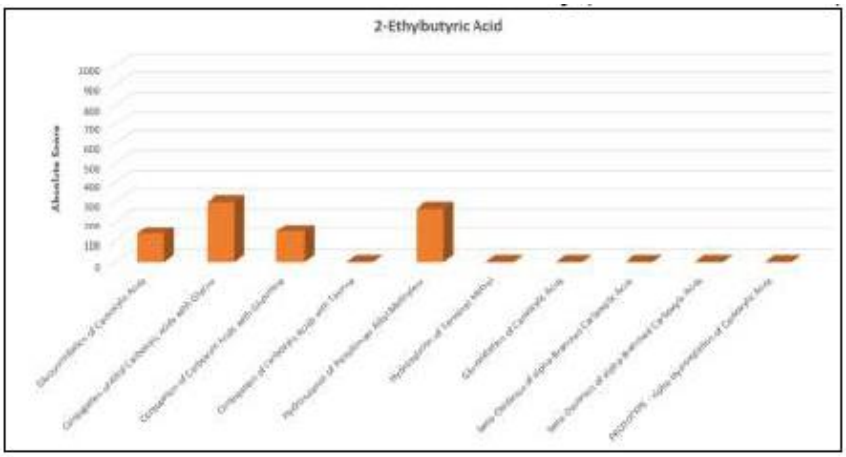


Predicted Metabolism of Case Study Target and Analogues

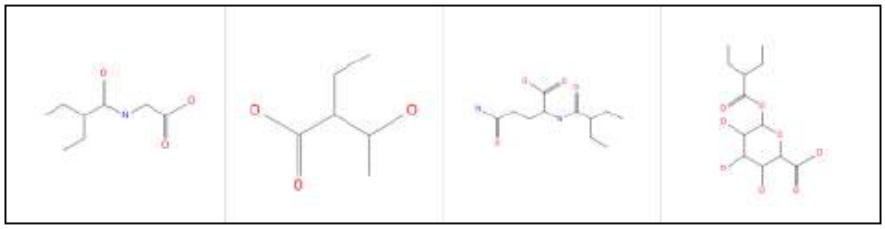
Target Compound: 2-Ethylbutyric Acid (CAS: 88-09-5)



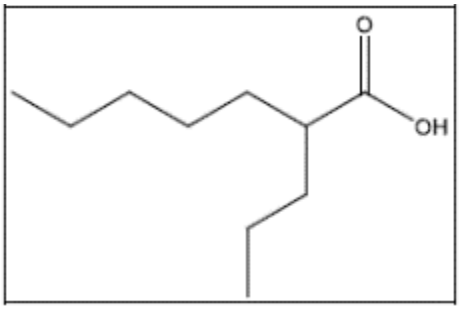
Predicted Metabolism at First Generation Only (see General Methods)



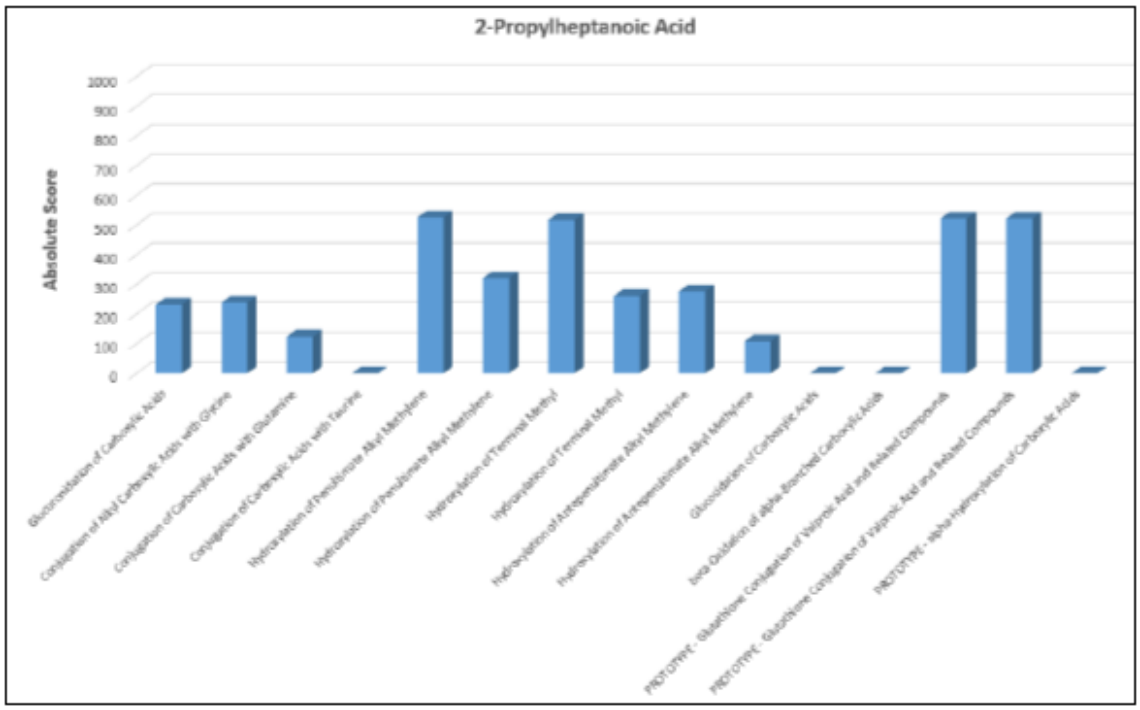
Metabolites Occurring as a Result of Biotransformations with Score > 0 (First Generation Only)



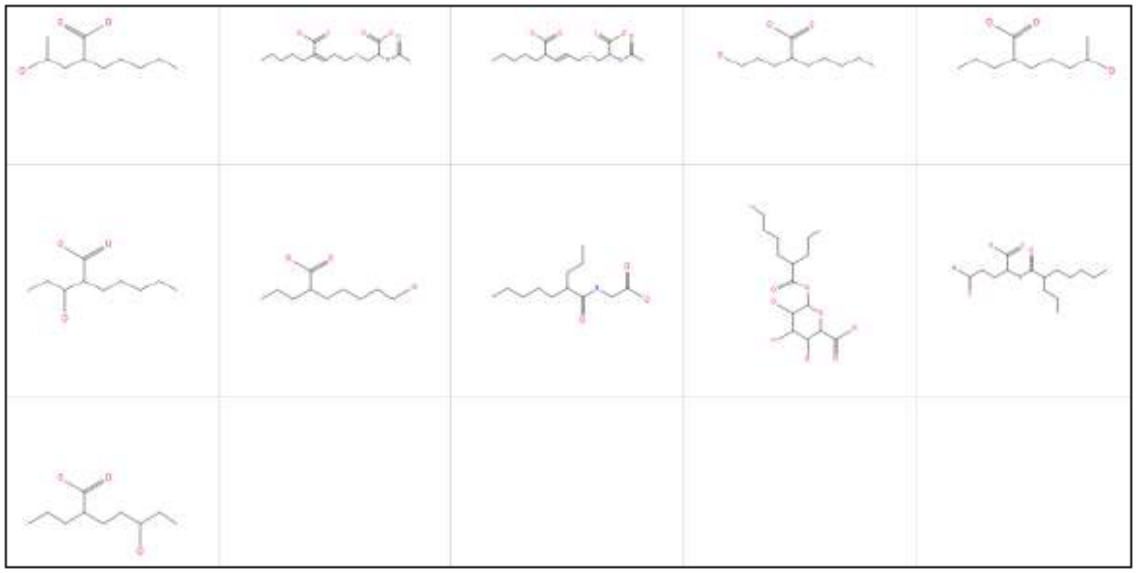
***Analogue Compound 1*: 2-Propylheptanoic Acid** *(CAS: 31080-39-4)*



Predicted Metabolism at First Generation Only (see General Methods)

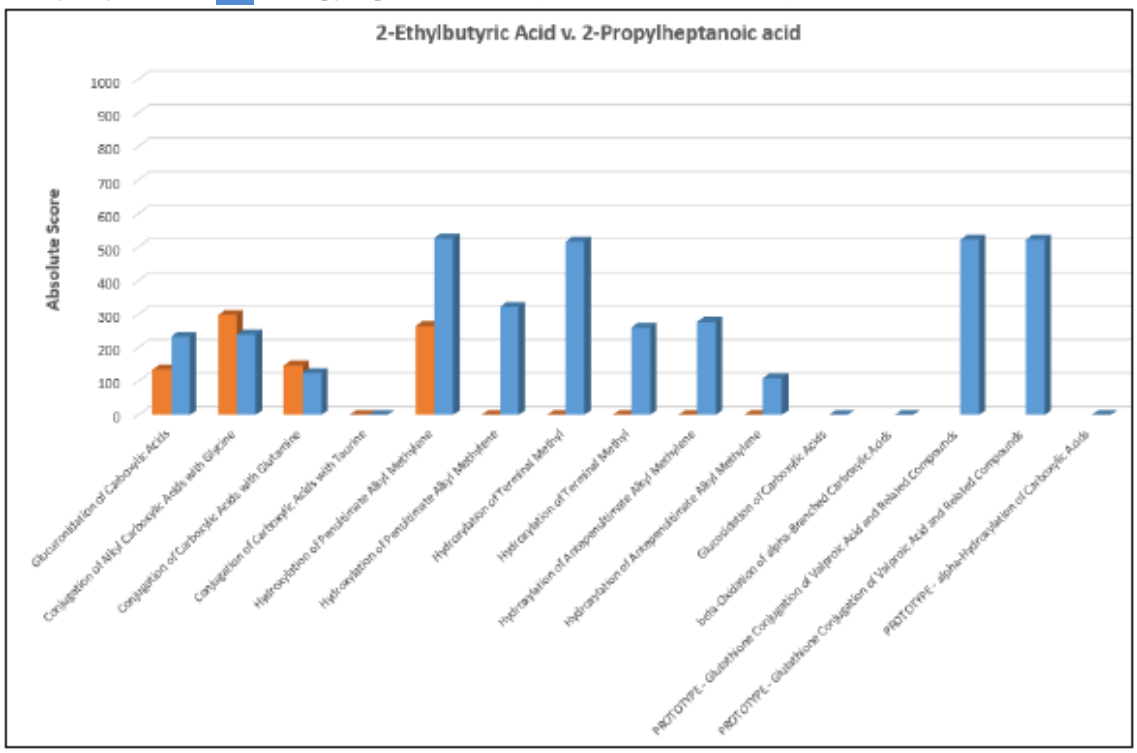


Metabolites Occurring as a Result of Biotransformations with Score > 0 (First Generation Only)



Metabolism Prediction: Comparison of Target Compound v. Analogue Compound 1 (First Generation Only)

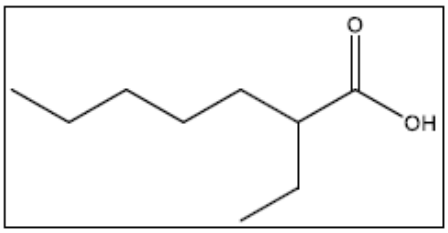
2-Ethylbutyric Acid 2-Propylheptanoic Acid (scores are not normalised)



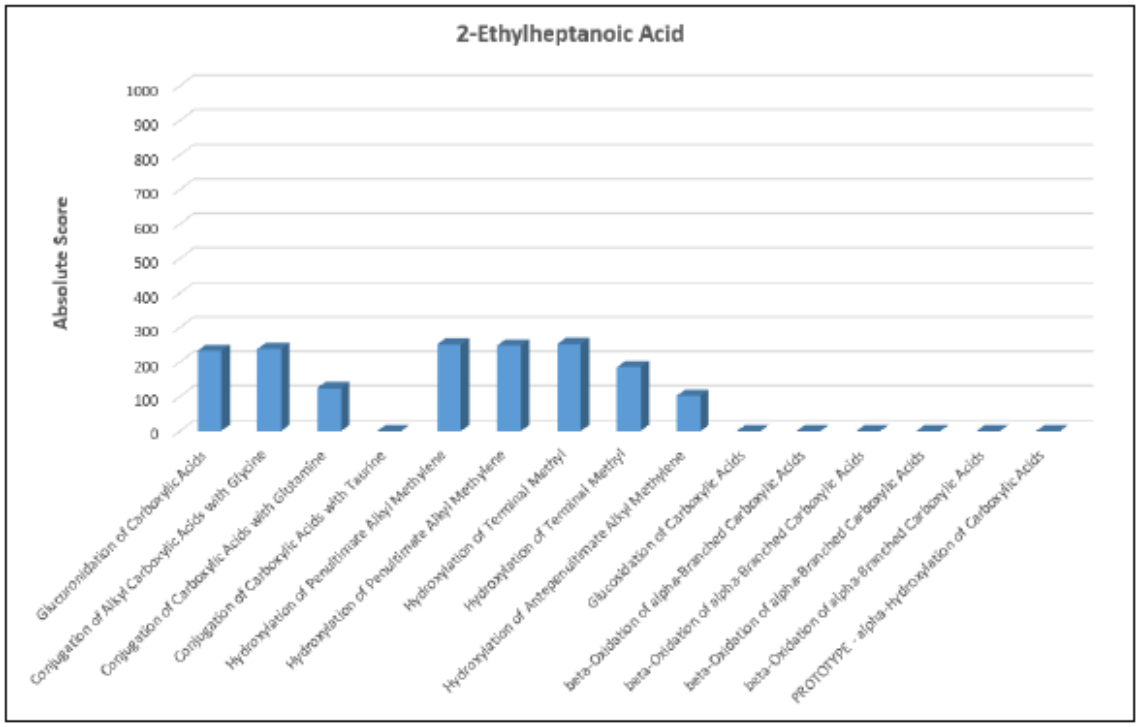
Similarity of First Generation Tree: Target Compound v. Analogue Compound 1 (See General Methods)

Biotransformation Fingerprint Method: 0.46

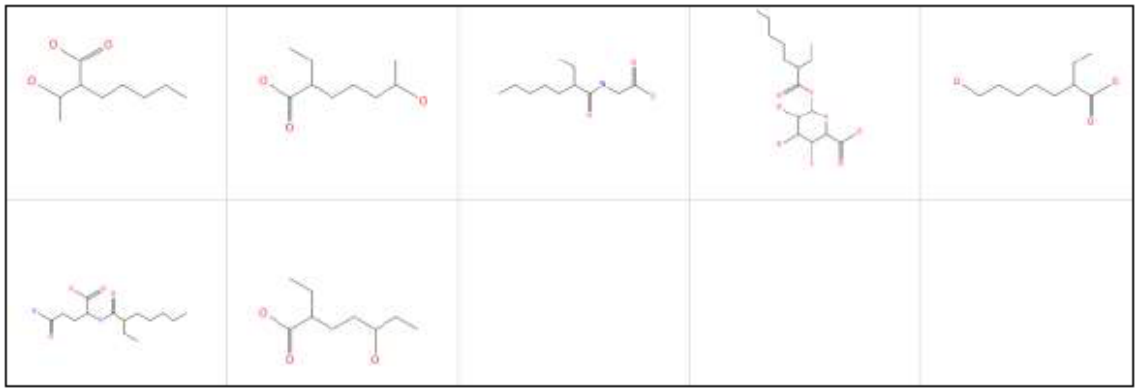
Analogue Compound 2: 2-Ethylheptanoic Acid *(CAS: 3274-29-1)*



Predicted Metabolism at First Generation Only (see General Methods)

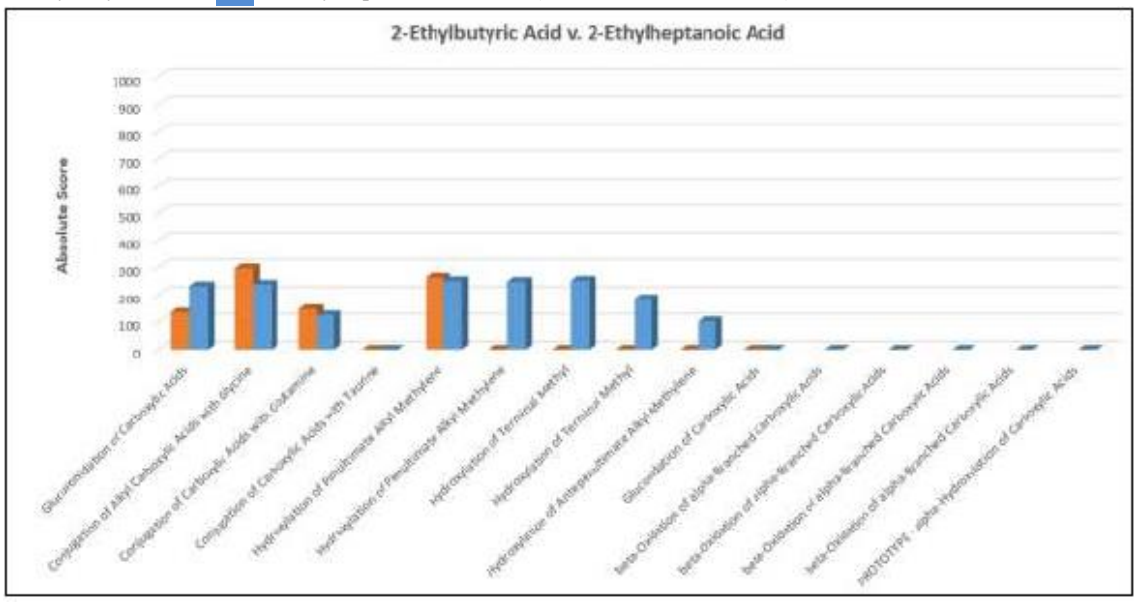


Metabolites Occurring as a Result of Biotransformations with Score > 0 (First Generation Only)



Metabolism Prediction: Comparison of Target Compound v. Analogue Compound 2 (First Generation Only)

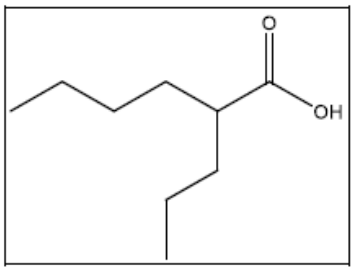
2-Ethylbutyric Acid 2-Ethylheptanoic Acid (scores are not normalised)



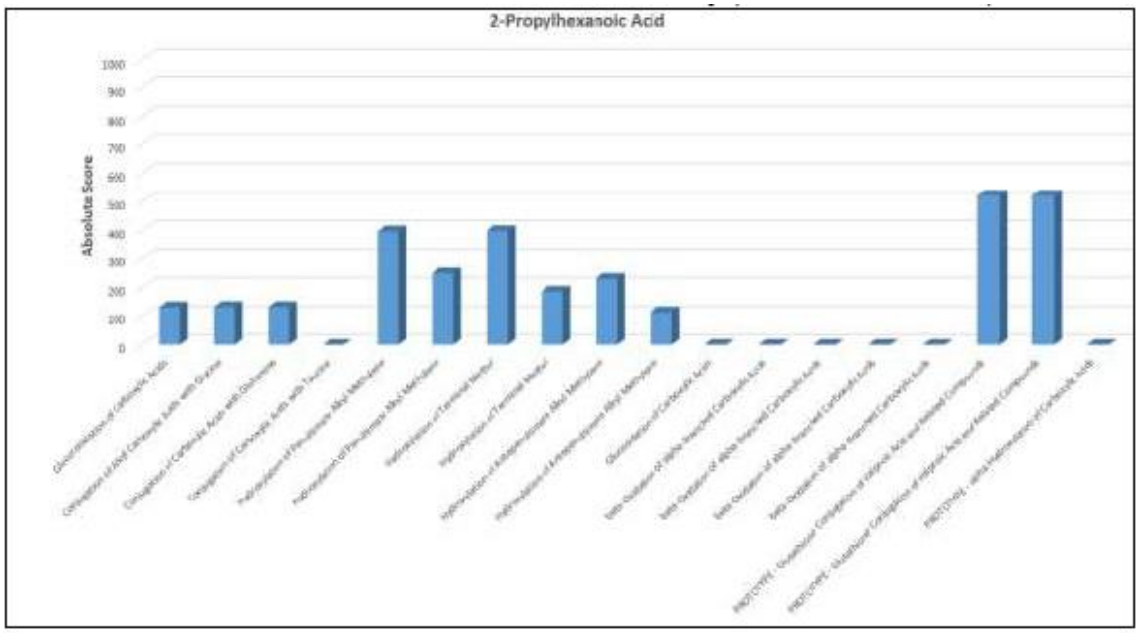
Similarity of First Generation Tree: Target Compound v. Analogue Compound 2 (See General Methods)

Biotransformation Fingerprint Method: 0.63

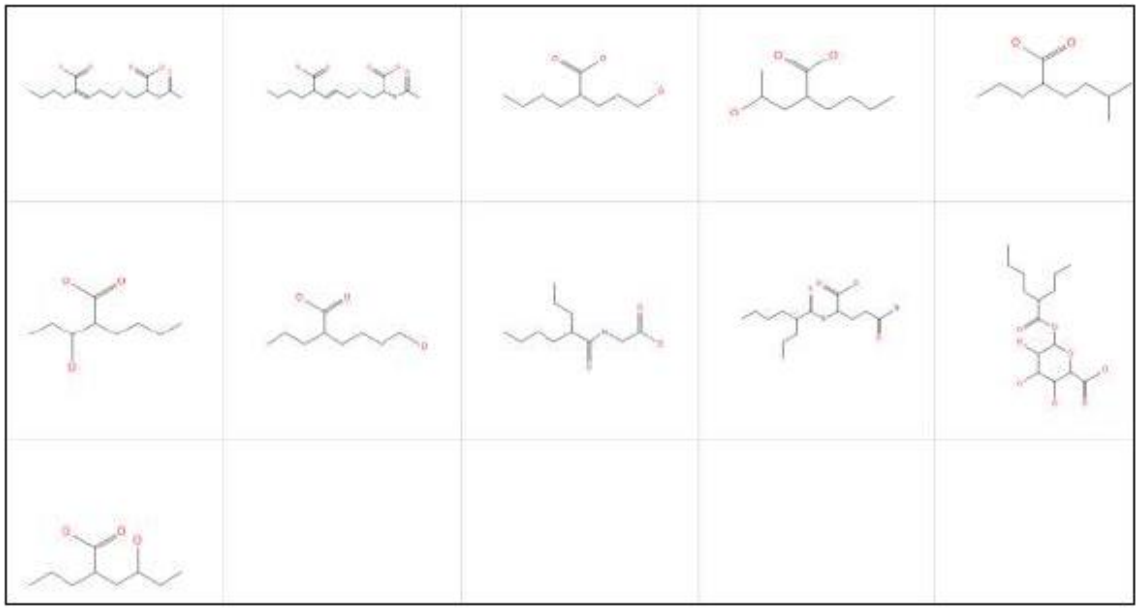
***Analogue Compound 3*: 2-Propylhexanoic Acid** *(CAS: 3274-28-0)*



Predicted Metabolism at First Generation Only (see General Methods)

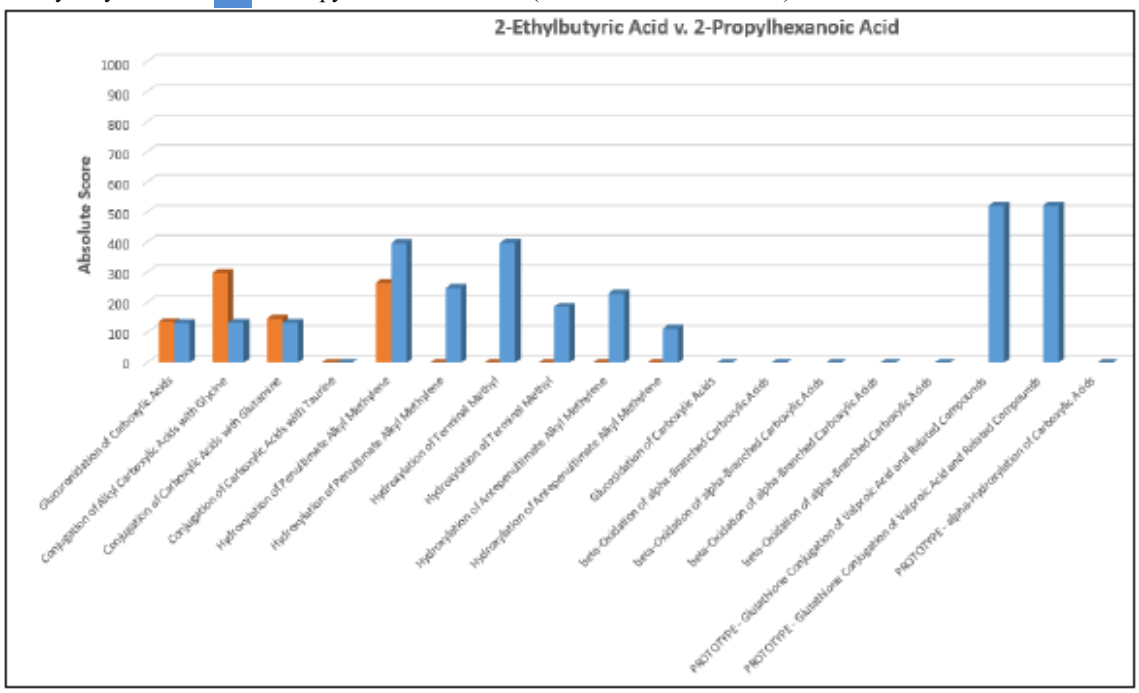


Metabolites Occurring as a Result of Biotransformations with Score > 0 (First Generation Only)



Metabolism Prediction: Comparison of Target Compound v. Analogue Compound 3 (First Generation Only)

2-Ethylbutyric Acid 2-Propylhexanoic Acid (scores are not normalised)

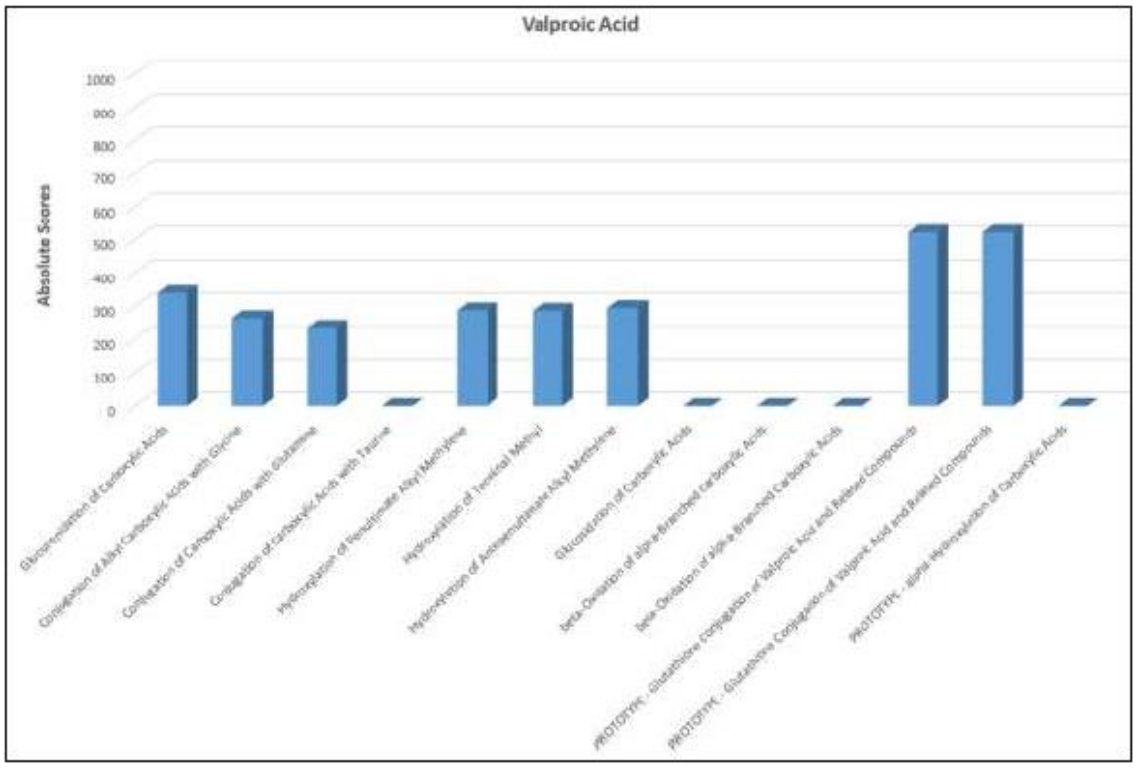


Similarity of First Generation Tree: Target Compound v. Analogue Compound 3 (See General Methods)

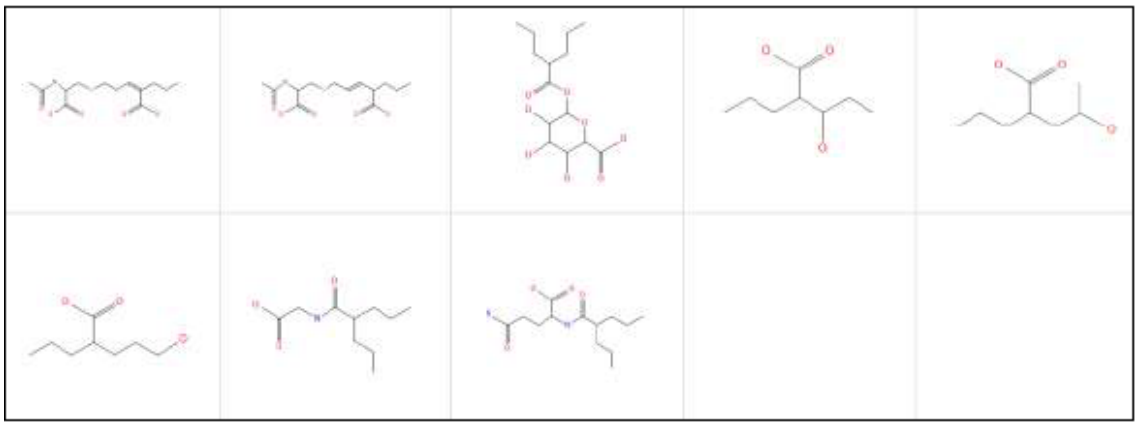
Biotransformation Fingerprint Method: 0.46

Analogue Compound 4: Valproic Acid *(CAS: 99-66-1)*

Predicted Metabolism at First Generation Only (see General Methods)

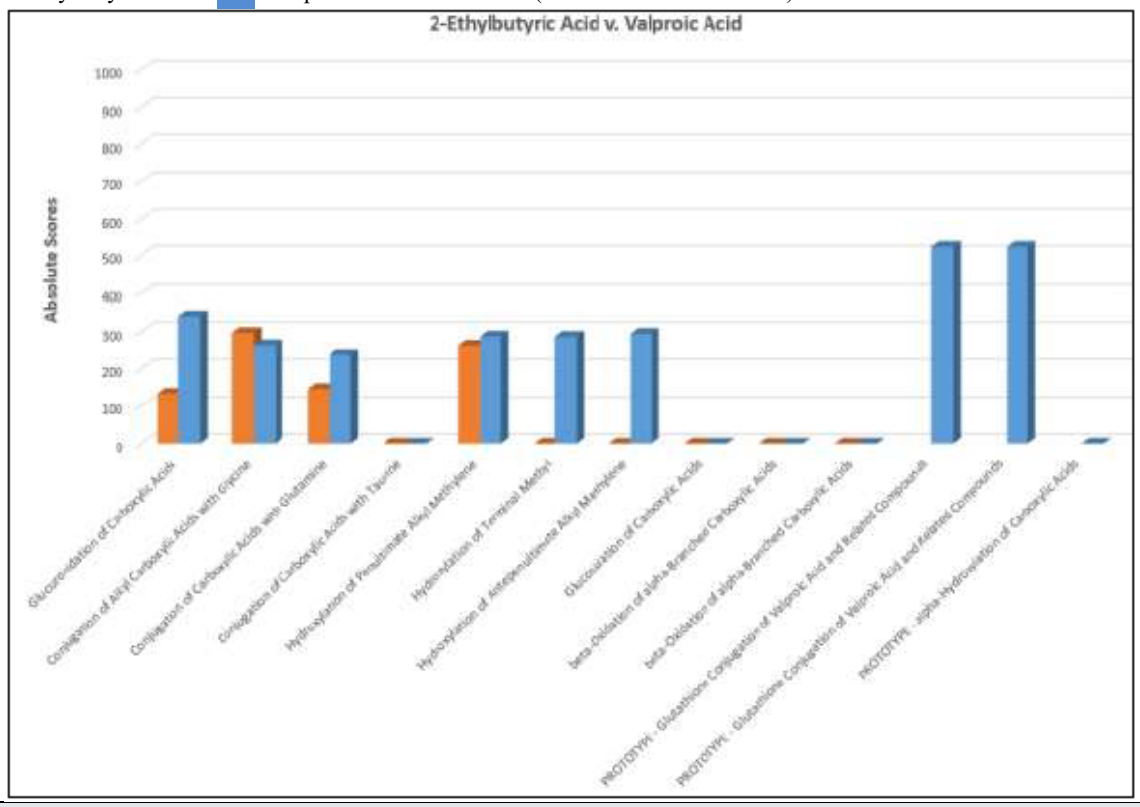


Metabolites Occurring as a Result of Biotransformations with Score > 0 (First Generation Only)



Metabolism Prediction: Comparison of Target Compound v. Analogue Compound 4 (First Generation Only)

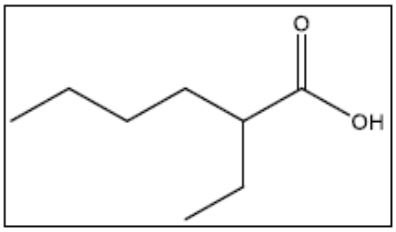
2-Ethylbutyric Acid Valproic Acid (scores are not normalised)



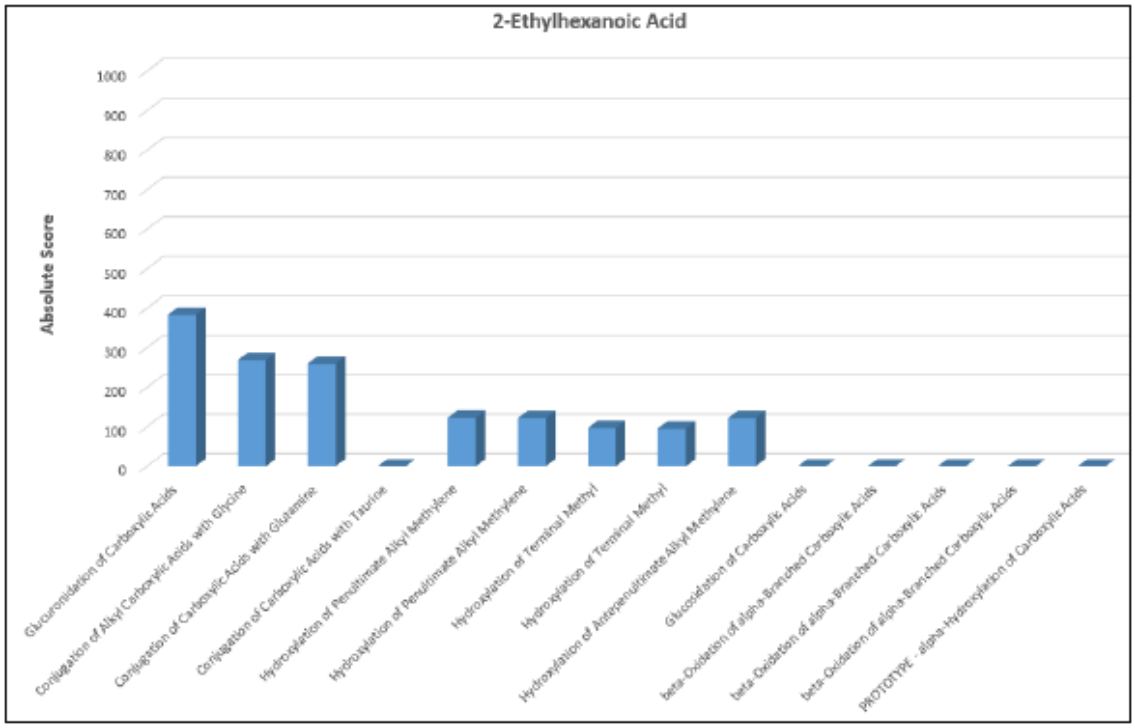
Similarity of First Generation Tree: Target Compound v. Analogue Compound 4 (See General Methods)

Biotransformation Fingerprint Method: 0.46

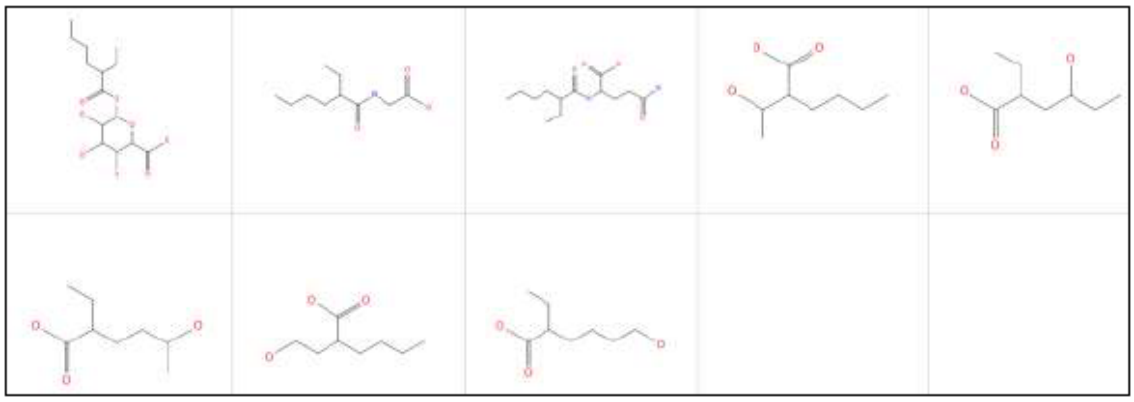
***Analogue Compound 5*: 2-Ethylhexanoic Acid** *(CAS: 149-57-5)*



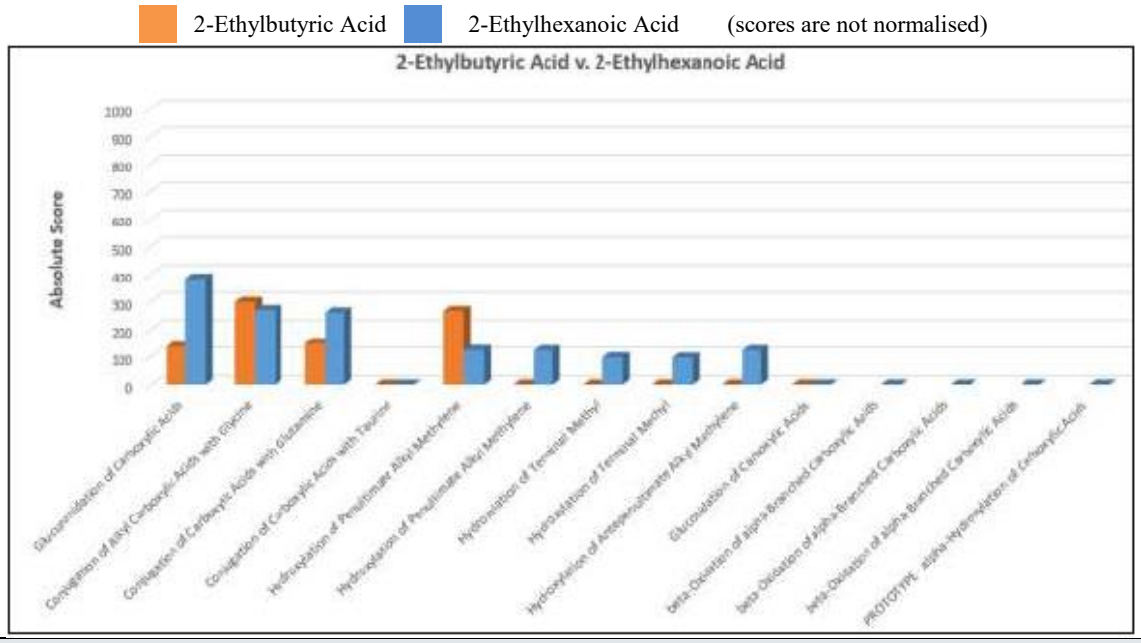
Predicted Metabolism at First Generation Only (see General Methods)



Metabolites Occurring as a Result of Biotransformations with Score > 0 (First Generation Only)



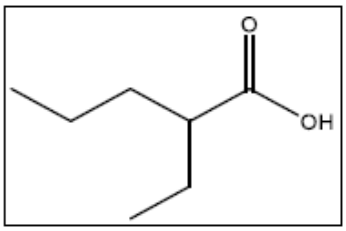
Metabolism Prediction: Comparison of Target Compound v. Analogue Compound 5 (First Generation Only)



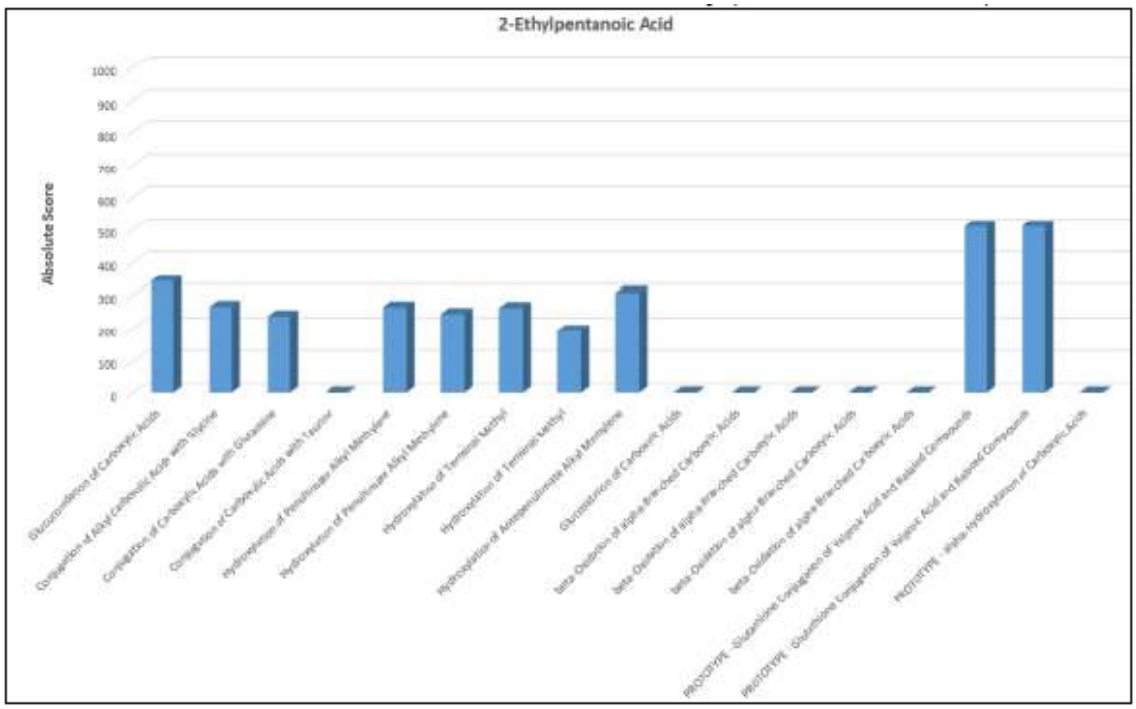
Similarity of First Generation Tree: Target Compound v. Analogue Compound 5 (See General Methods)

Biotransformation Fingerprint Method: 0.63

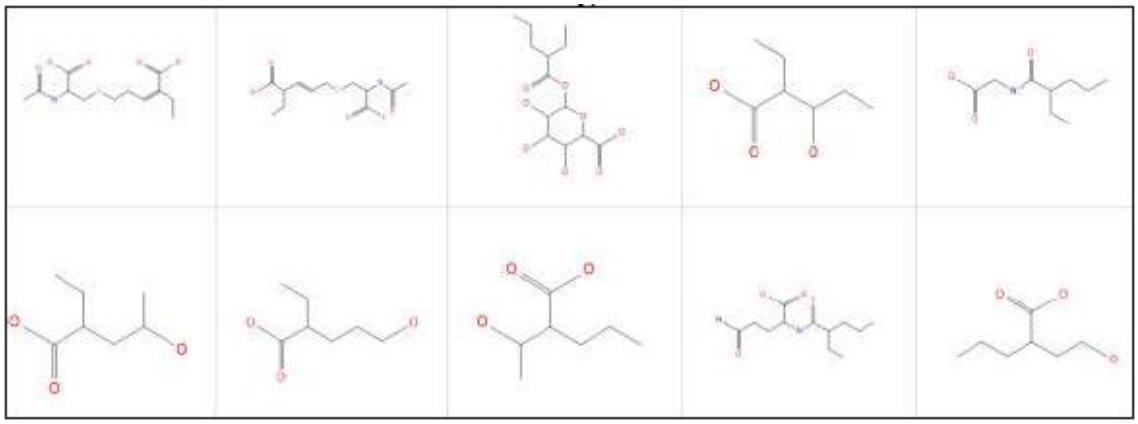
***Analogue Compound 6*: 2-Ethylpentanoic Acid** *(CAS: 20225-24-5)*



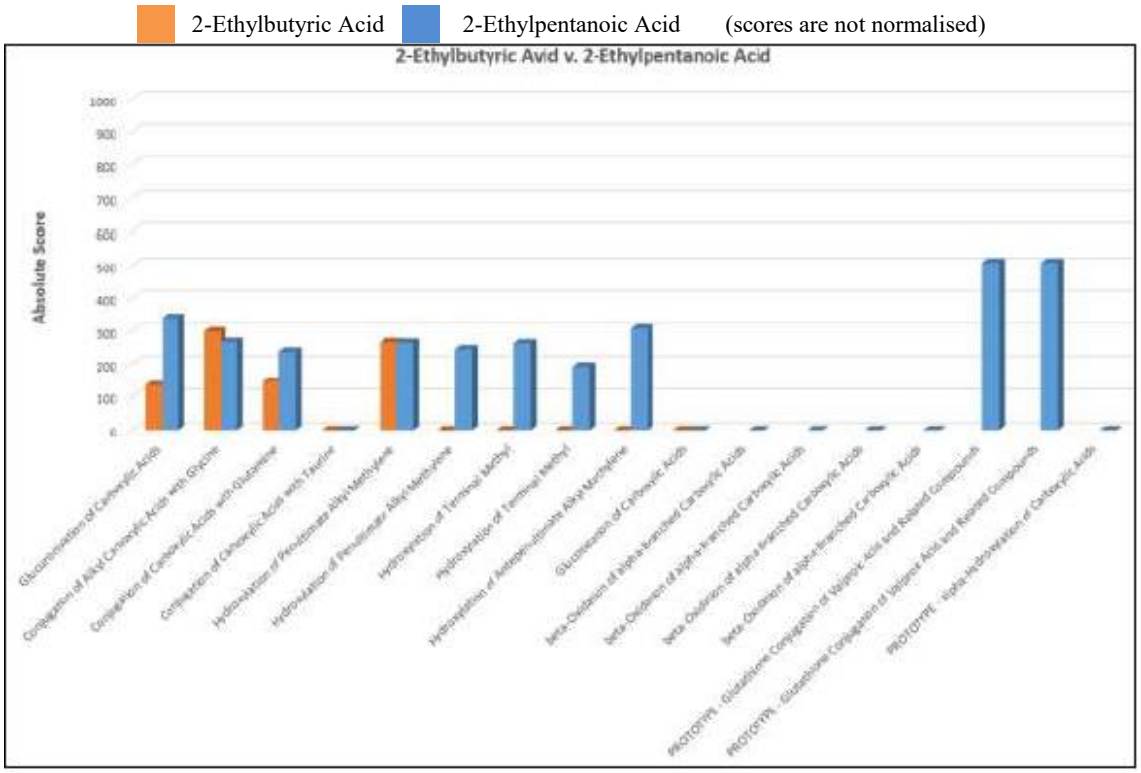
Predicted Metabolism at First Generation Only (see General Methods)



Metabolites Occurring as a Result of Biotransformations with Score > 0 (First Generation Only)



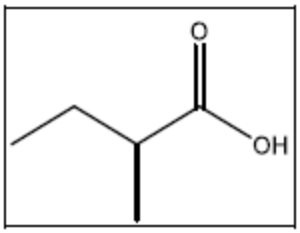
Metabolism Prediction: Comparison of Target Compound v. Analogue Compound 6 (First Generation Only)



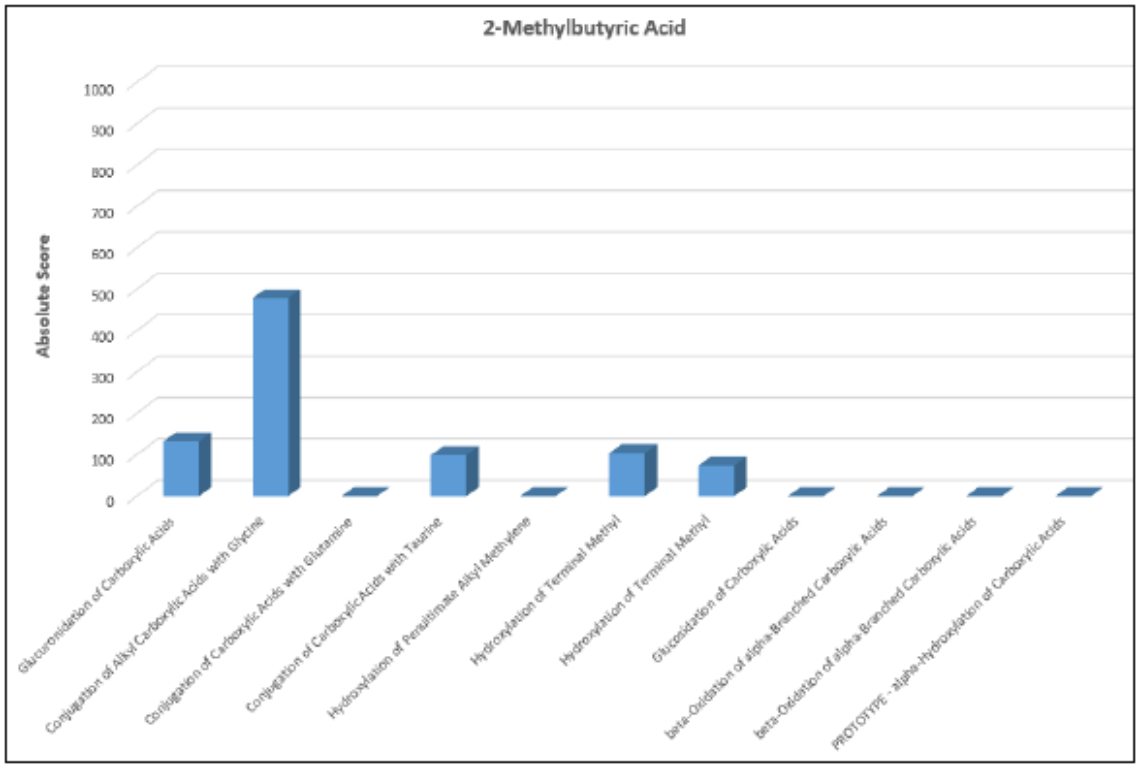
Similarity of First Generation Tree: Target Compound v. Analogue Compound 6 (See General Methods)

Biotransformation Fingerprint Method: 0.50

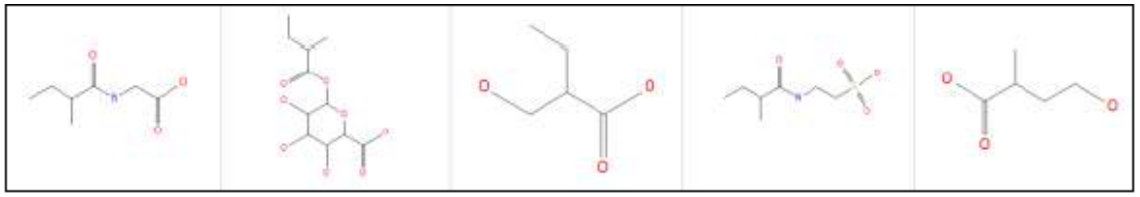
***Analogue Compound 7*: 2-Methylbutyric Acid** *(CAS: 1730-91-2)*



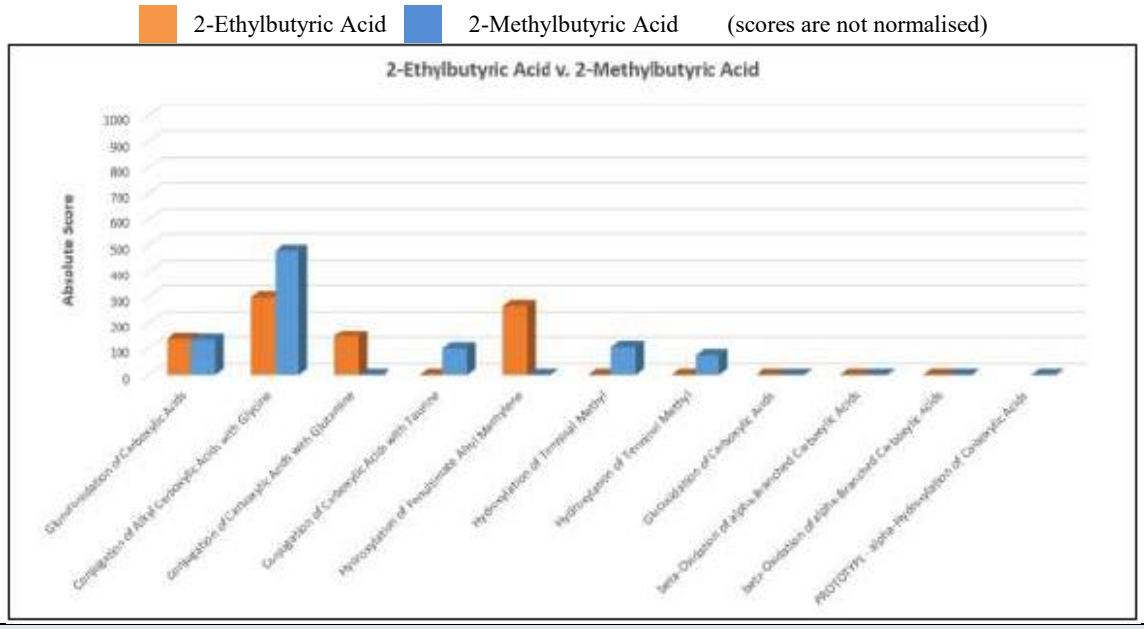
Predicted Metabolism at First Generation Only (see General Methods)



Metabolites Occurring as a Result of Biotransformations with Score > 0 (First Generation Only)



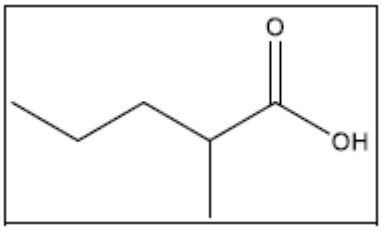
Metabolism Prediction: Comparison of Target Compound v. Analogue Compound 7 (First Generation Only)



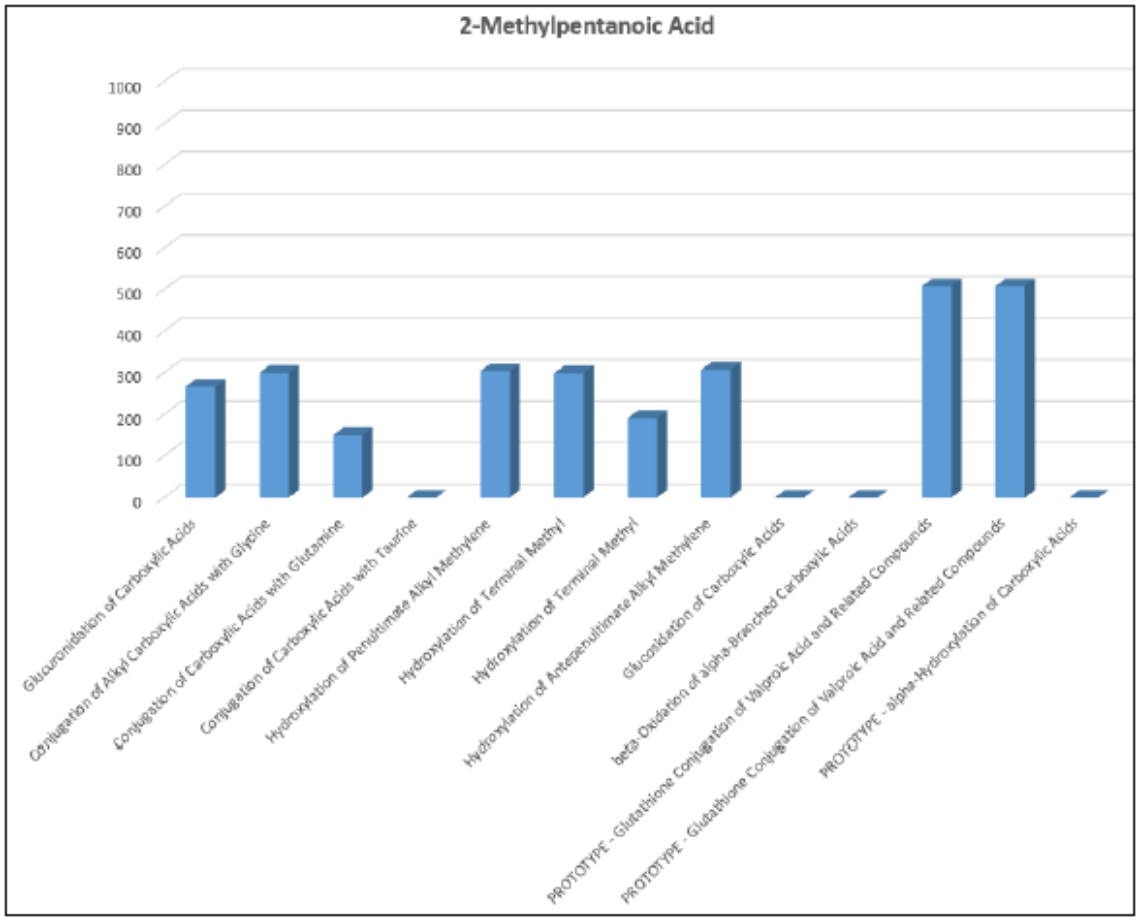
Similarity of First Generation Tree: Target Compound v. Analogue Compound 7 (See General Methods)

Biotransformation Fingerprint Method: 0.67

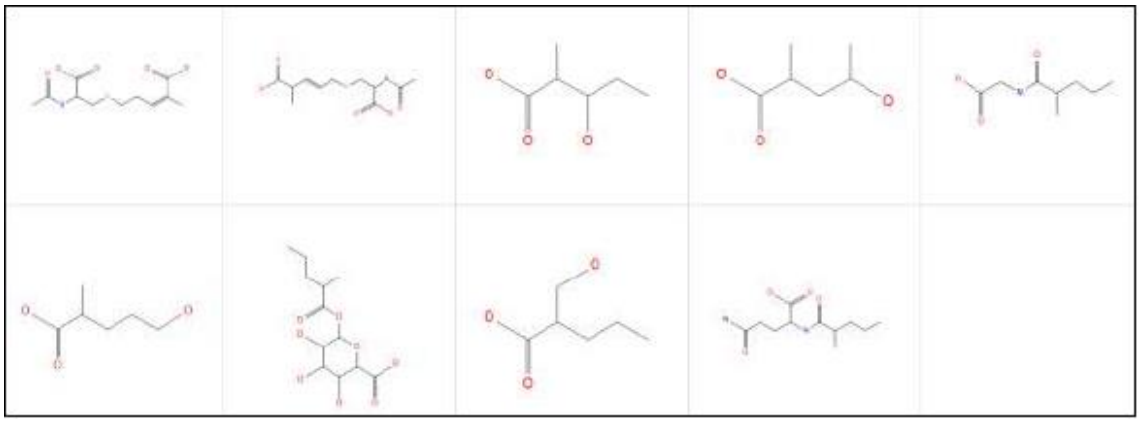
***Analogue Compound 8*: 2-Methylpentanoic Acid** *(CAS: 97-61-0)*



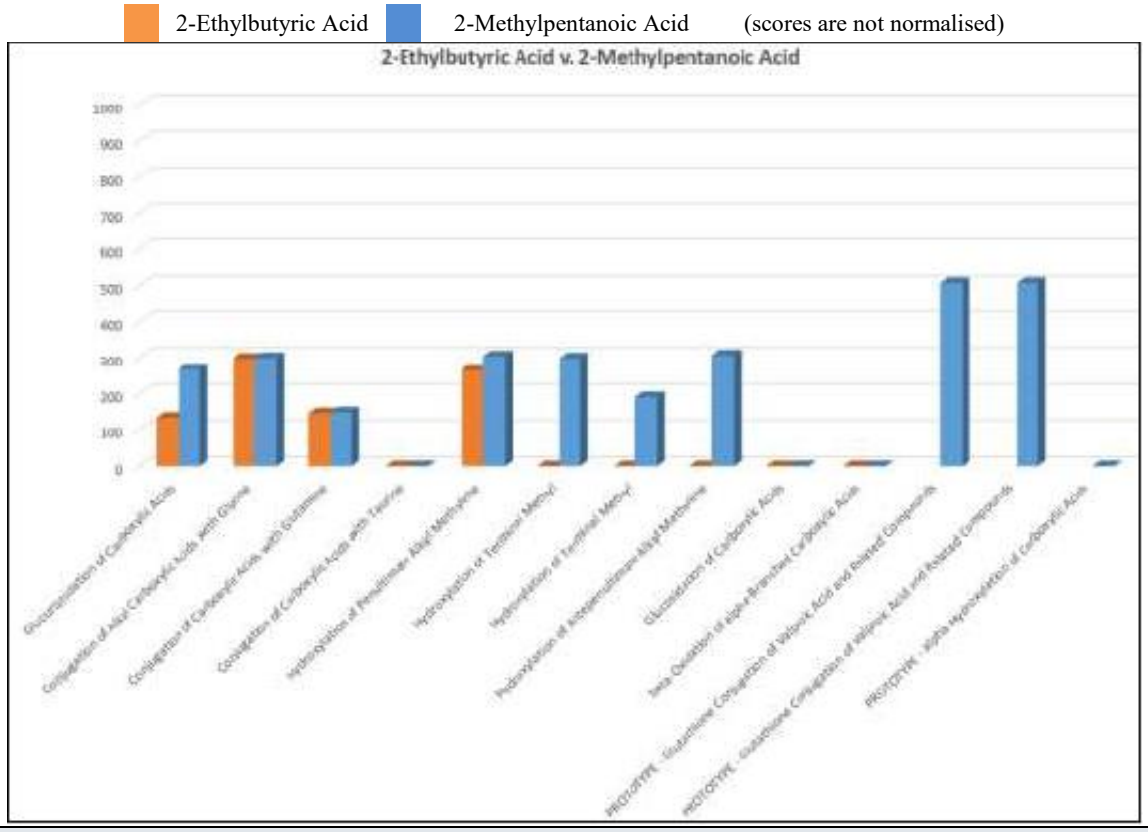
Predicted Metabolism at First Generation Only (see General Methods)



Metabolites Occurring as a Result of Biotransformations with Score > 0 (First Generation Only)



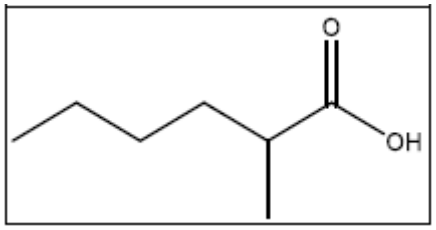
Metabolism Prediction: Comparison of Target Compound v. Analogue Compound 8 (First Generation Only)



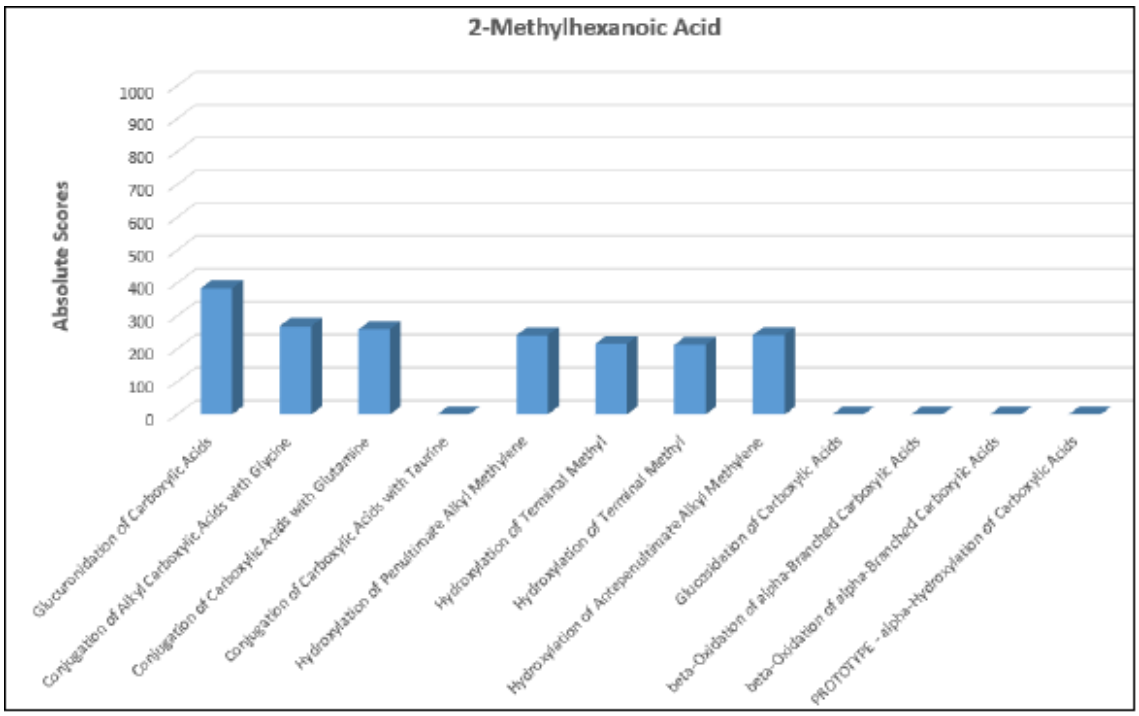
Similarity of First Generation Tree: Target Compound v. Analogue Compound 8 (See General Methods)

Biotransformation Fingerprint Method: 0.44

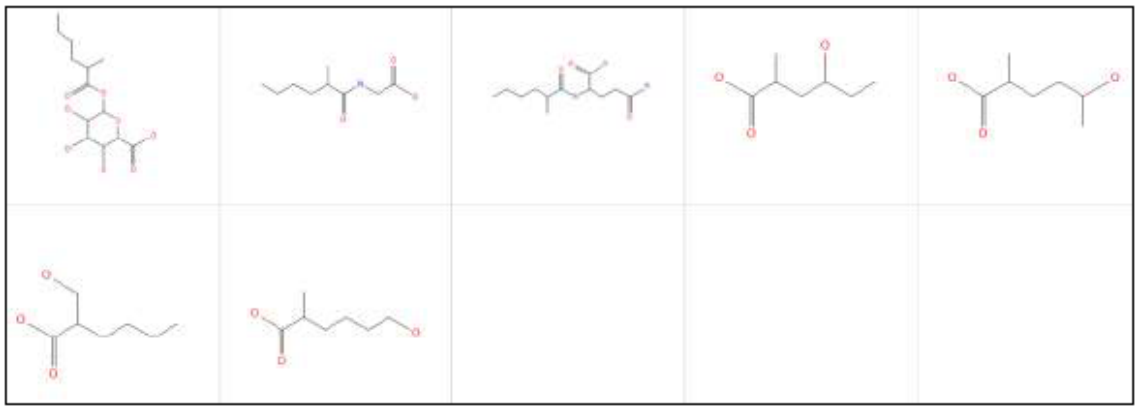
***Analogue Compound 9*: 2-Methylhexanoic Acid** *(CAS: 4536-23-6)*



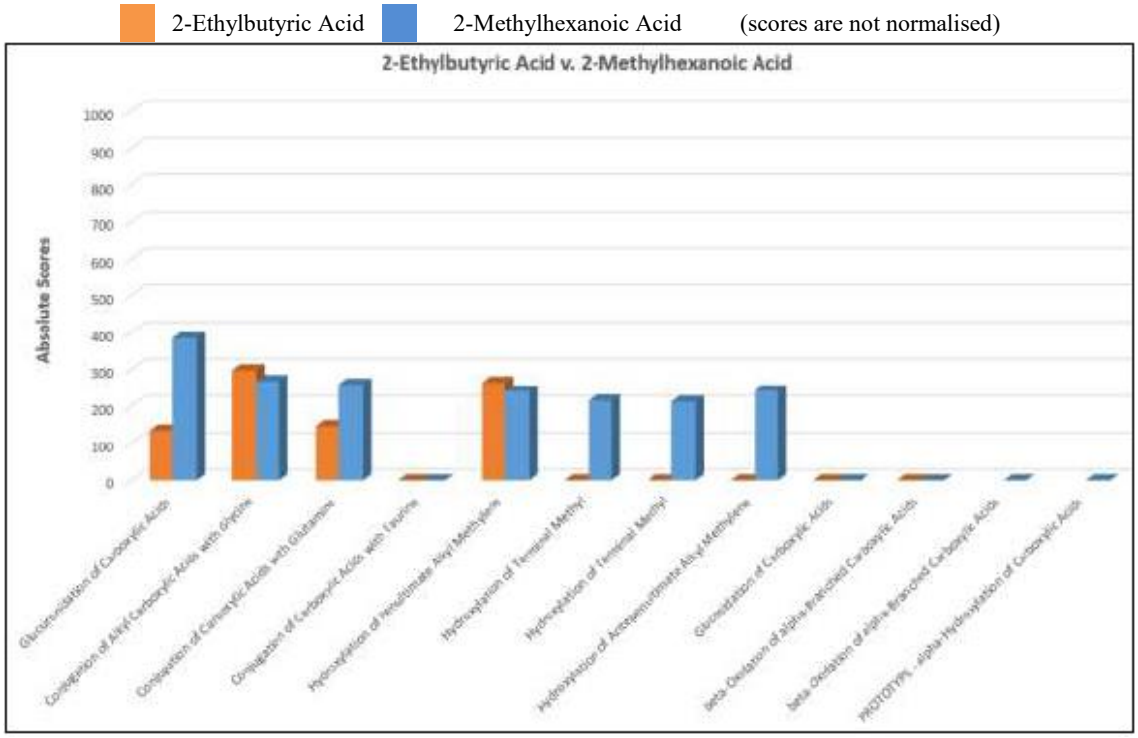
Predicted Metabolism at First Generation Only (see General Methods)



Metabolites Occurring as a Result of Biotransformations with Score > 0 (First Generation Only)



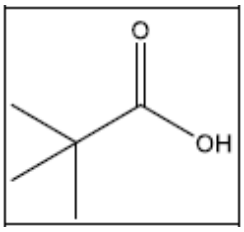
Metabolism Prediction: Comparison of Target Compound v. Analogue Compound 9 (First Generation Only)



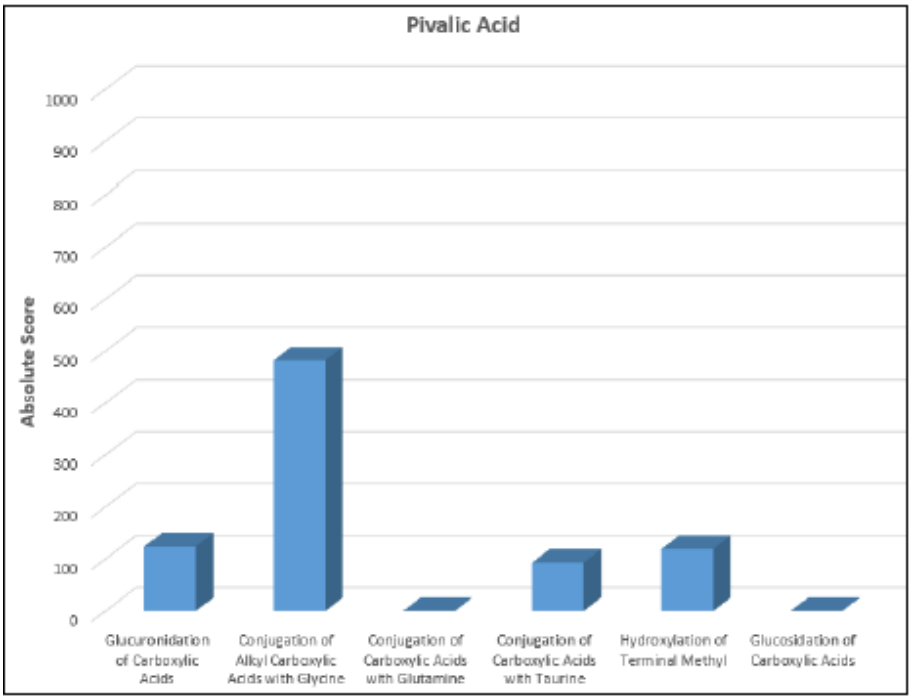
Similarity of First Generation Tree: Target Compound v. Analogue Compound 9 (See General Methods)

Biotransformation Fingerprint Method: 0.50

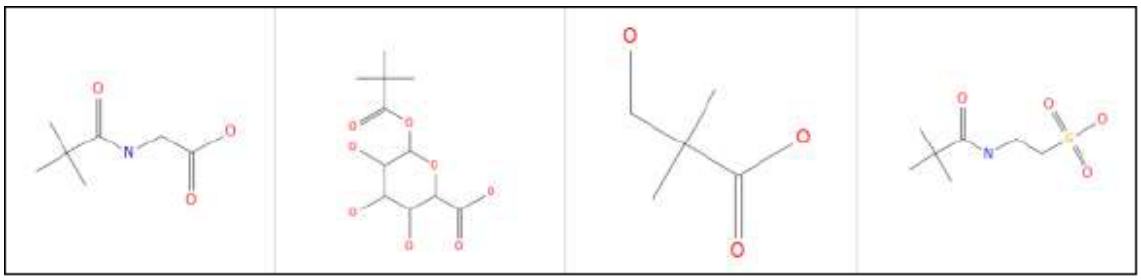
***Analogue Compound 10*: Pivalic Acid** *(CAS: 75-98-9)*



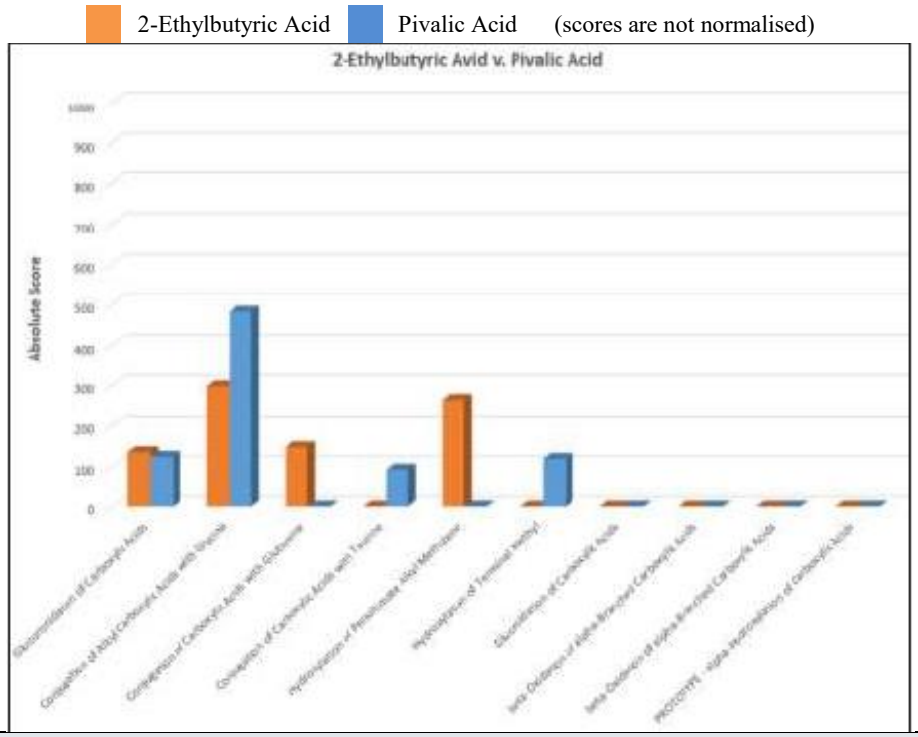
Predicted Metabolism at First Generation Only (see General Methods)



Metabolites Occurring as a Result of Biotransformations with Score > 0 (First Generation Only)



Metabolism Prediction: Comparison of Target Compound v. Analogue Compound 10 (First Generation Only)

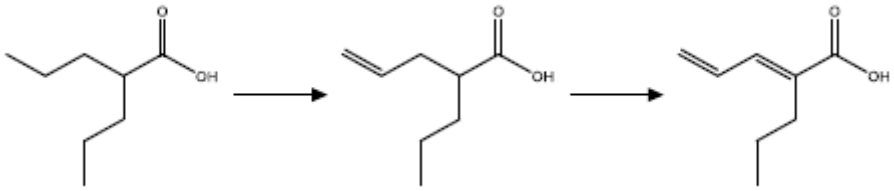


Similarity of First Generation Tree: Target Compound v. Analogue Compound 10 (See General Methods)

Biotransformation Fingerprint Method: 0.25

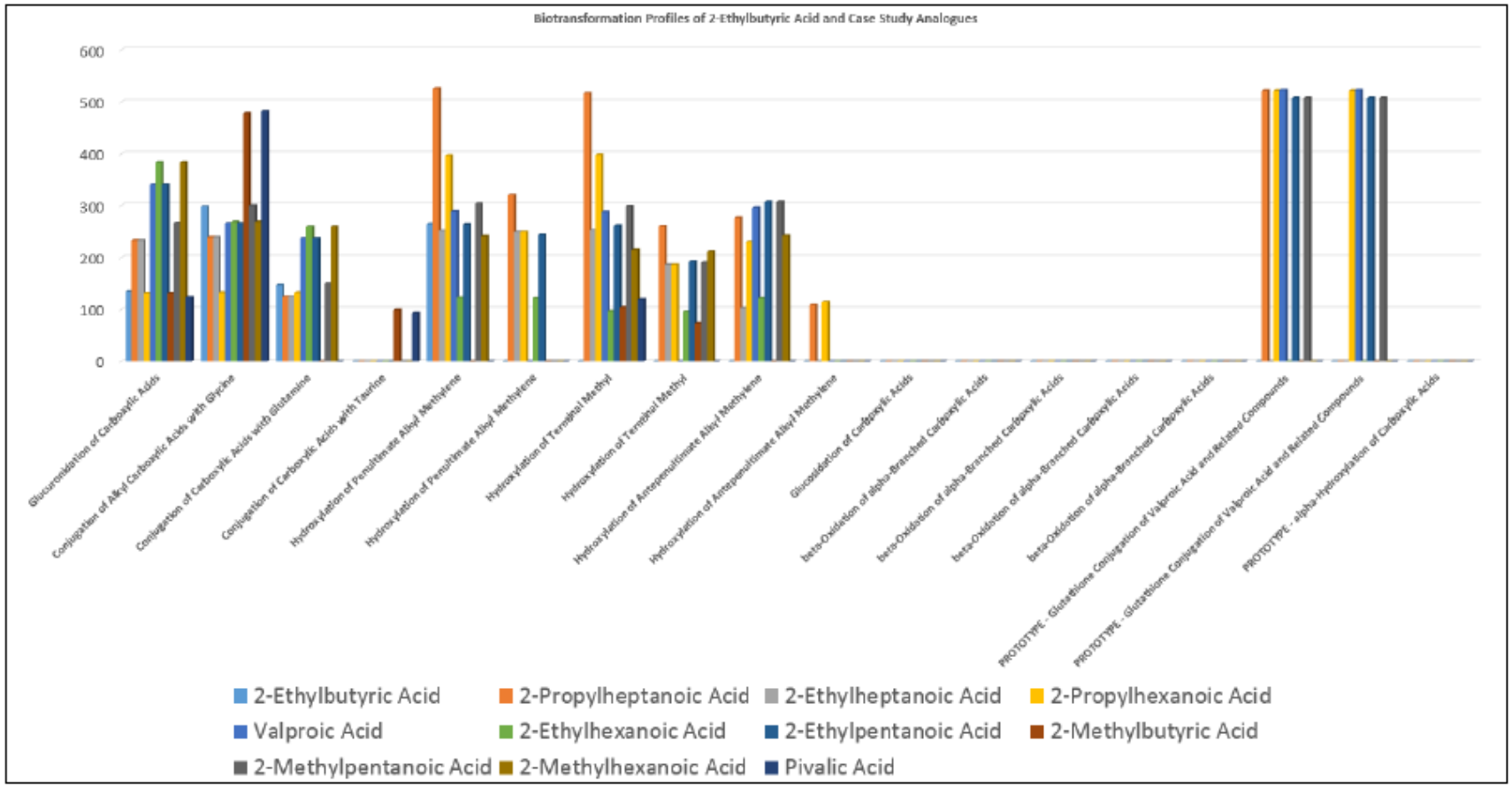
General Comments on Prediction and Comparison to Published Literature

The most extensively studied and metabolically characterised analogue in this study is valproic acid itself and this has an extensive and complex metabolic profile as already discussed. Meteor Nexus is able to predict all phase II biotransformations with the exception of the adenylate (adenosine monophosphate) conjugate reported in a couple of studies. This is generally not a significant or detectable conjugate in the study of the vast majority of compounds with a carboxylic acid function. Whilst betaoxidation is predicted for most analogues, the score assigned to the biotransformations is always zero. The reason for this is that the Meteor Nexus choice of nearest neighbour algorithm requires the metabolite structure in the data set, in the case of beta-oxidation (and in respect of the Meteor Nexus definition of this biotransformation) the final product that results after the final thiolytic cleavage and CoA deconjugation. These steps has been hinted at in the literature but never observed for valproic acid or any of its analogues. The theoretical structures and intermediates are available though for inspection in the Meteor Nexus trees. For substrates with a branched propyl chain, Meteor Nexus predicts conjugation with glutathione (biotransformation 506). Within this pathway, there are intermediates which arise as a result of sequential CYP-oxidation (terminal dehydrogenation) and beta-oxidation (internal dehydrogenation):

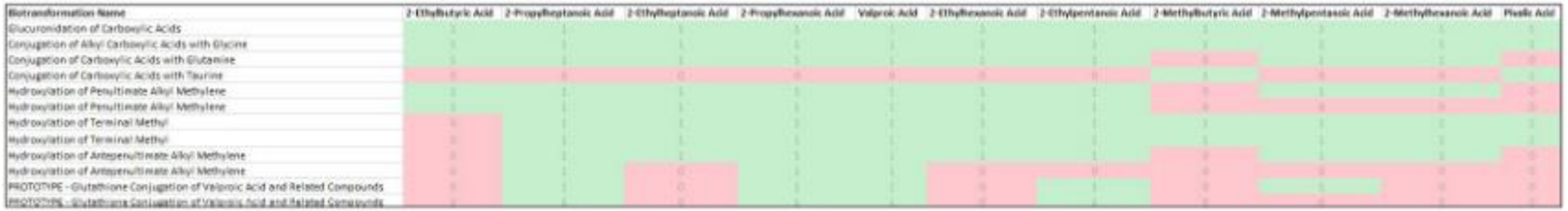


which is followed by 1,6-conjugate addition of glutathione to the 2-propyl-penta-2,4-dienoic acid. This reaction has been observed only for valproic, 4-ene- and 2,4-diene valproic acids but would in any case seem to be peculiar to propyl groups as a terminal dehydrogenation is involved. All of the major CYP oxidations seen in the literature are predicted by Meteor Nexus including some of the non-obvious metabolites such as those resulting from lactone formation, however it is necessary to examine the larger, multi-generational trees in order to observe the predictions of these downstream metabolites. Understanding the limitation of the zero score for beta-oxidations, Meteor Nexus predicts well for this class in respect of the major and most often observed metabolites. The general (and expected) trend in prediction (and limited observation) is that the relative amount of carboaliphatic oxidation increases as the length of the alkyl chain and the side branch increases. Some degree of conjugation (by a combination of glucuronic acid, glycine, glutamine and taurine) is always predicted and these are usually the most significant contributors to clearance of these compounds *in vivo*.

Combined Graph of Biotransformation Scores for Target Compound and Analogue Compounds



ransformation Fingerprints (Heatmap) for Target and Analogue Compounds



Tanimoto Similarity Matrix Based on Biotransformation Fingerprints for Target and Analogue Compounds

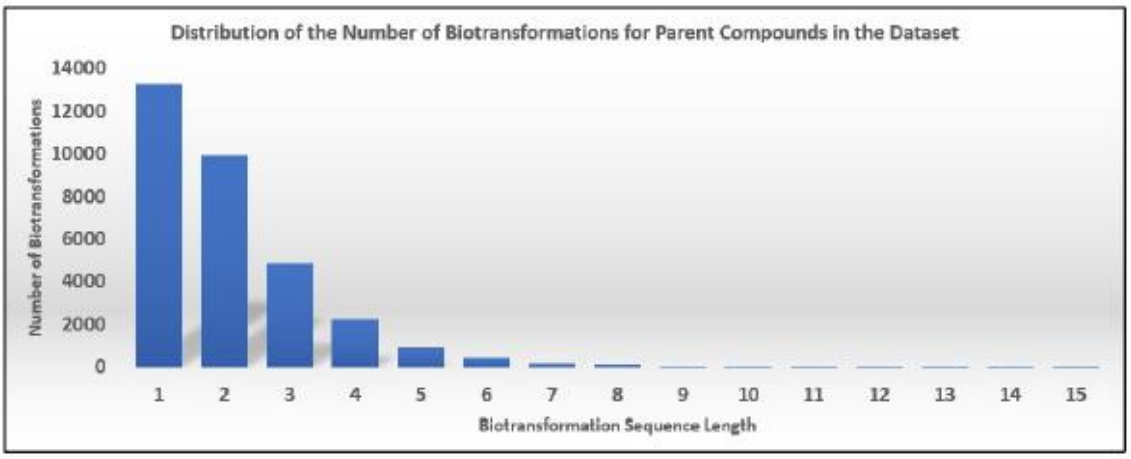
	2-Ethylbutyric Acid	2-Propylheptanoic Acid	2-Ethylheptanoic Acid	2-Propylhexanoic Acid	Valproic Acid	2-Ethylhexanoic Acid	2-Ethylpentanoic Acid	2-Methylbutyric Acid	2-Methylpentanoic Acid	2-Methylhexanoic Acid	Pivalic Acid
2-Ethylbutyric Acid	1	0.46	0.63	0.46	0.46	0.63	0.5	0.67	0.44	0.5	0.25
2-Propylheptanoic Acid	0.46	1	0.73	1	1	0.73	0.91	0.33	0.73	0.64	0.33
2-Ethylheptanoic Acid	0.63	0.73	1	0.73	0.73	1	0.8	0.44	0.78	0.87	0.44
2-Propylhexanoic Acid	0.46	1	0.73	1	1	0.73	0.9	0.33	0.73	0.64	0.33
Valproic Acid	0.46	1	0.73	1	1	0.73	0.9	0.33	0.73	0.63	0.33
2-Ethylhexanoic Acid	0.63	0.73	1	0.73	0.73	1	0.8	0.44	0.78	0.86	0.44
2-Ethylpentanoic Acid	0.5	0.91	0.8	0.9	0.9	0.8	1	0.36	0.8	0.7	0.5
2-Methylbutyric Acid	0.67	0.33	0.44	0.33	0.33	0.44	0.36	1	0.44	0.5	1
2-Methylpentanoic Acid	0.44	0.73	0.78	0.73	0.73	0.78	0.8	0.44	1	0.88	0.44
2-Methylhexanoic Acid	0.5	0.64	0.87	0.64	0.63	0.86	0.7	0.5	0.88	1	0.5
Pivalic Acid	0.25	0.33	0.44	0.33	0.33	0.44	0.5	1	0.44	0.5	1

Tanimoto Coefficient Code:

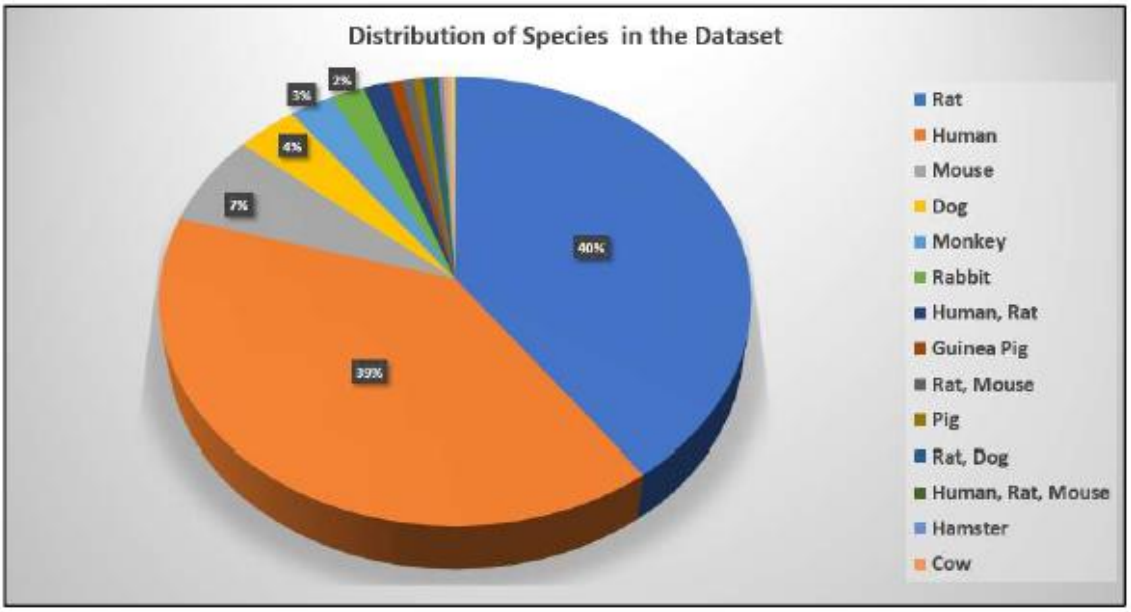
0
0.01-0.33
0.34-0.66
0.67-0.99
1

Description of the Meteor NEXUS Model: Essential Statistics* Regarding the Lhasa Limited Metabolism Dataset and Definition of Applicability Domain

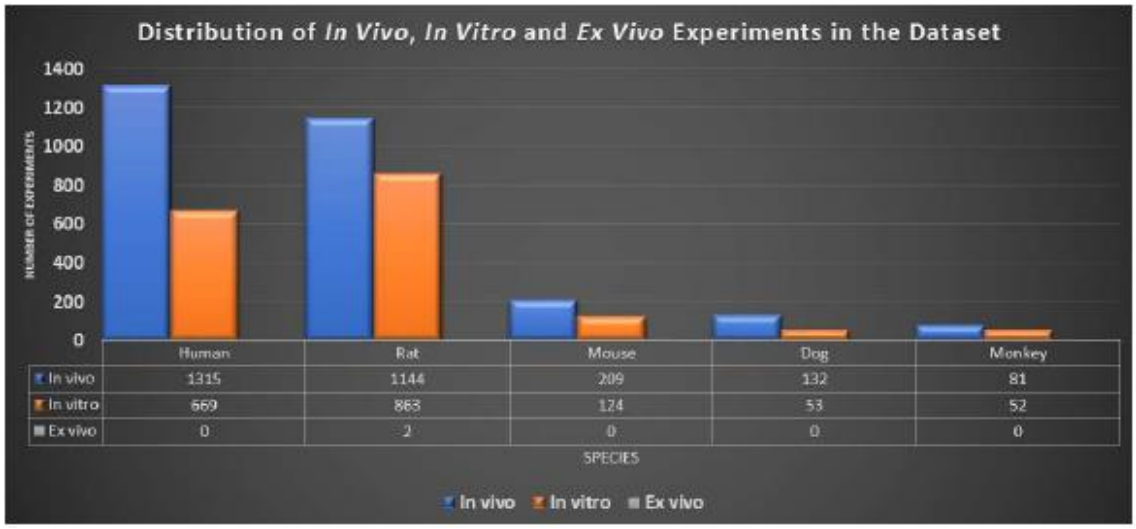
Number of Unique Parent Compounds:	2739
Number of Metabolic Biotransformations:	31897



Experiment Type	Number of Experiments
In vitro:	2764
In vivo	2477
Ex Vivo	2



*Numbers refer to Lhasa Limited Metabolism Dataset version 2.2.0 used in this study.
Valproic acid and pivalic acid in our mock submission are part of the dataset.



Analysis of the Applicability Domain

In this section, the applicability domain is analysed. The applicability domain indicates whether or not the model/database used for the prediction of, for example, biotransformation scores cover compounds with comparable features to the predicted compounds. Applicability domains can be analysed in many different ways20. Here we report how well the source and target compounds are covered with regard to their structural and physico-chemical properties in the Meteor Nexus knowledge base.

Meteor Nexus is an knowledge based system. Its rules have been derived by human experts over many years from study and knowledge of the metabolic chemist ry literature and as such are not built on the dataset in the same way as a QSAR or some other model is mathematically built on a training set. However, biotransformations scores, which give us a level of confidence in the occurrence of a particular biotransformation, are predicated16 on a set of nearest neighbour parent compounds drawn from the dataset. In that respect the metabolism dataset is a viable "surrogate", training set for a definition of the applicability domain in this analysis.

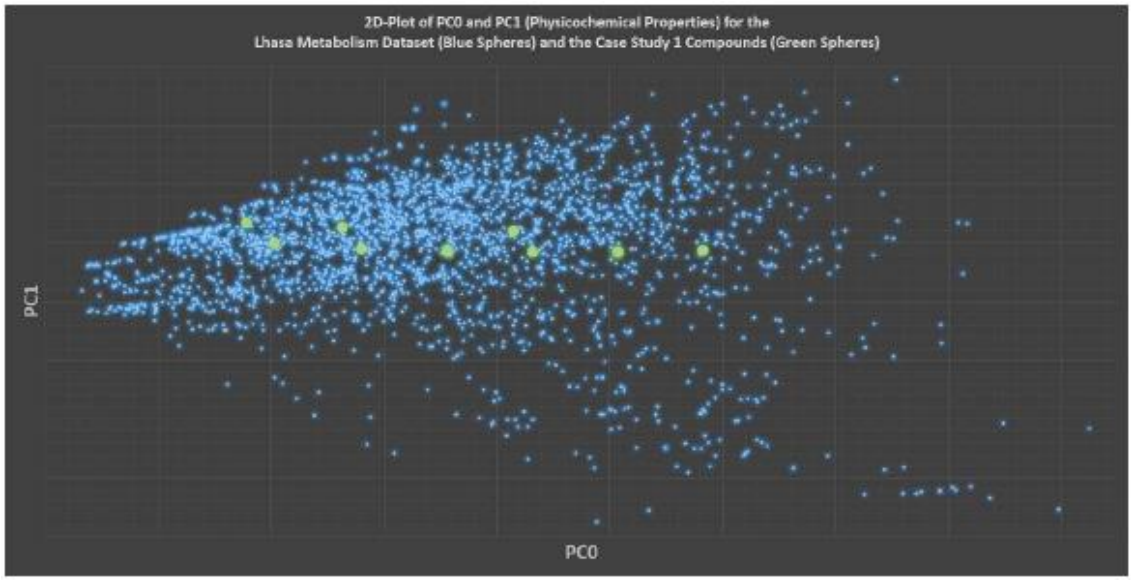
Principal Component Analysis and Molecular Fingerprint Definition of the Lhasa Limited Metabolism Dataset and the Mock Submission Target and Analogue Compounds

Principal Component Analysis (PCA) is a mathematical method for dimensionality reduction that allows for multidimensional datasets to be visualised using two- or three-dimensional plots with minimal loss of information[10, 11]. In this case, we chose 32 structural and physicochemical properties of two datasets namely the Lhasa Limited Metabolism Dataset and the chosen case study 1 target and analogue compounds. Each compound in the dataset is represented by a 32-dimensional vector defined by the physicochemical properties calculated using RDKit and CDK nodes in KNIME.

Calculated Descriptors from RDKit and CDK
SlogP, SMR, LabuteASA, TPSA, AMW, NumLipinskiHBA, NumLipinskiHBD, NumRotatableBonds, NumHBD, NumHBA, NumAmideBonds, NumHeteroAtoms, NumHeavyAtoms, NumAtoms, NumStereocenters, NumUnspecifiedStereocenters, NumRings, NumAromaticRings, NumSaturatedRings, NumAliphaticRings, NumAromaticHeterocycles, NumSaturatedHeterocycles, NumAliphaticHeterocycles, NumAromaticCarbocycles, NumSaturatedCarbocycles, NumAliphaticCarbocycles, FractionCSP3, Atomic Polarizabilities, Bond Polarizabilities, VABC Volume Descriptor, Largest Pi Chain, Molecular Weight.

Each value was normalised using 'z-scores', meaning that each descriptor value is transformed such that the value in each column has a mean of 0.0 and a standard deviation of 1.0. Each chemical in the dataset can also be represented as a molecular fingerprint12 using RDKit.

The first principal component calculated (PC0) accounts for as much of the variability in the original (normalised) data as possible, with each succeeding component (PC1, PC2 etc.), which are projected orthogonally to each other, accounting for as much of the remaining variability as possible. To visualise the chemical space occupied by the case study parent structures, the points calculated for the Lhasa Dataset (small blue spheres) were plotted using the values for the first two principal components (PC0 and PC1) and the case study parent compounds points were overlaid (large green spheres). This is shown in the following figure:

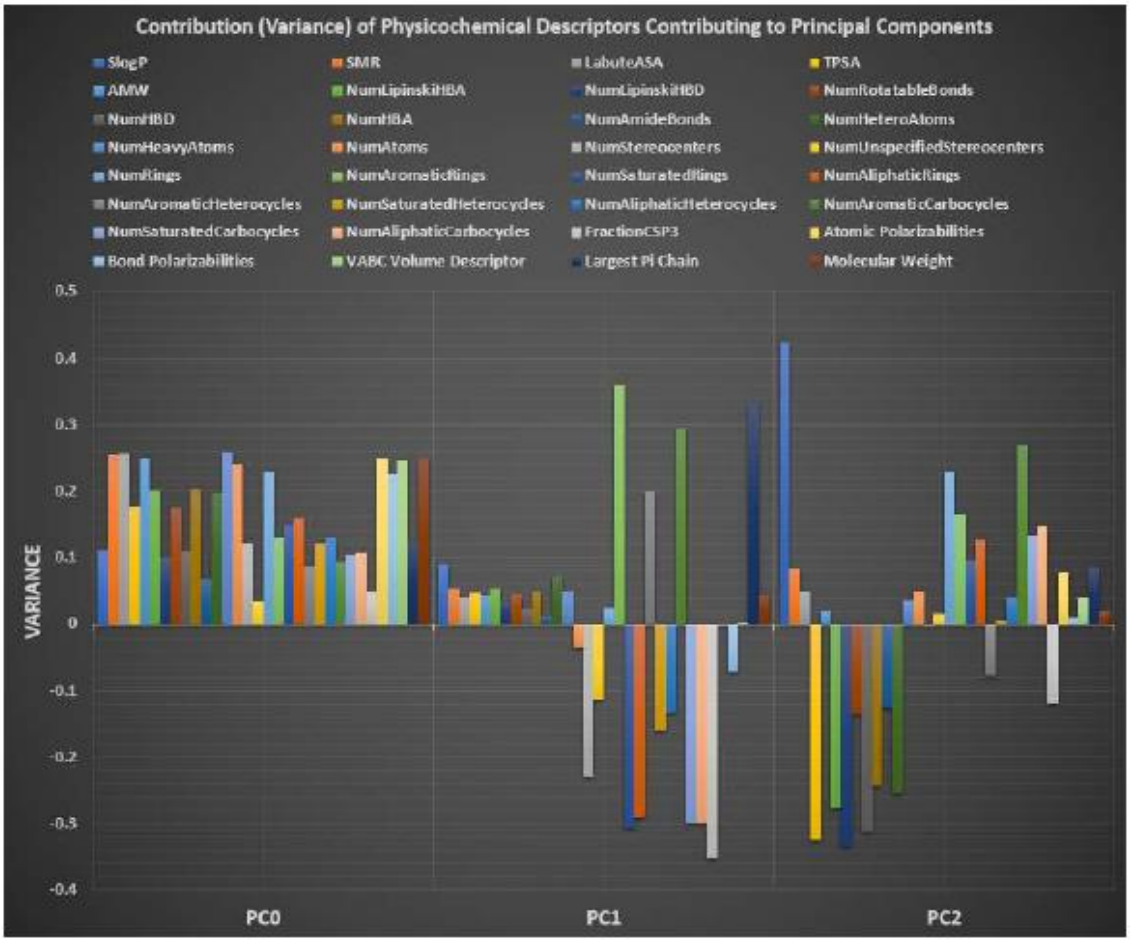


The figure shows that the case study compounds lie within the domain of the chemical space described by the Lhasa Metabolism Dataset as defined by these principal components.

If the data is projected using a third principal component (PC2) then the overall 3D space occupied by the case study compounds can be shown to lie within the chemical space defined by the minimum and maximum values for each axis of the three principal components as shown in the following table:

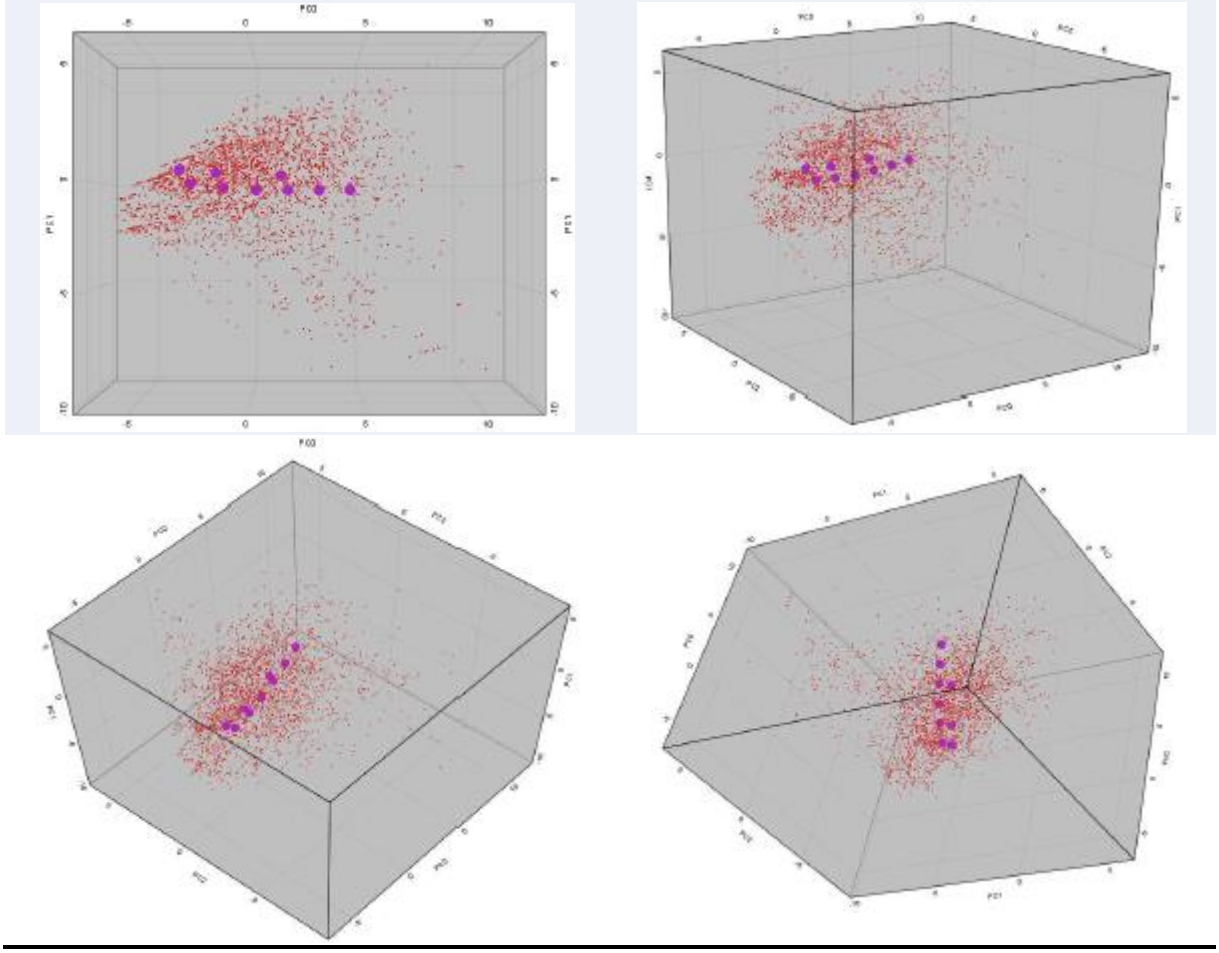
	PC0	PC1	PC2
Minimum Value of Principal Component	**-6.3509**	**-9.4763**	**-8.8195**
2,2-dimethylpropanoic acid	-3.4332	0.6506	-0.729
2-methylbutanoic acid	-2.9244	-0.0348	-0.83
2-ethylbutanoic acid	-1.728	0.5106	-0.5607
2-methylpentanoic acid	-1.3918	-0.2201	-0.5276
2-methylhexanoic acid	0.127	-0.3091	-0.1925
2-ethylpentanoic acid	0.127	-0.3091	-0.1925
Valproic acid	1.3011	0.3928	0.1298
2-ethylhexanoic acid	1.6372	-0.3379	0.163
2-ethylheptanoic acid	3.1417	-0.3265	0.5321
2-propylhexanoic acid	3.1417	-0.3265	0.5321
2-propylheptanoic acid	4.6422	-0.2871	0.9107
Maximum Value of Principal Component	**11.4787**	**5.5558**	**5.8887**

Contribution (variance) of physicochemical descriptors to each principal component is shown in the following chart:



Visualisation of three principal components is best achieved interactively - in this instance a KNIME 2D/3D Scatterplot node was employed. Some representative visualisations are shown below:

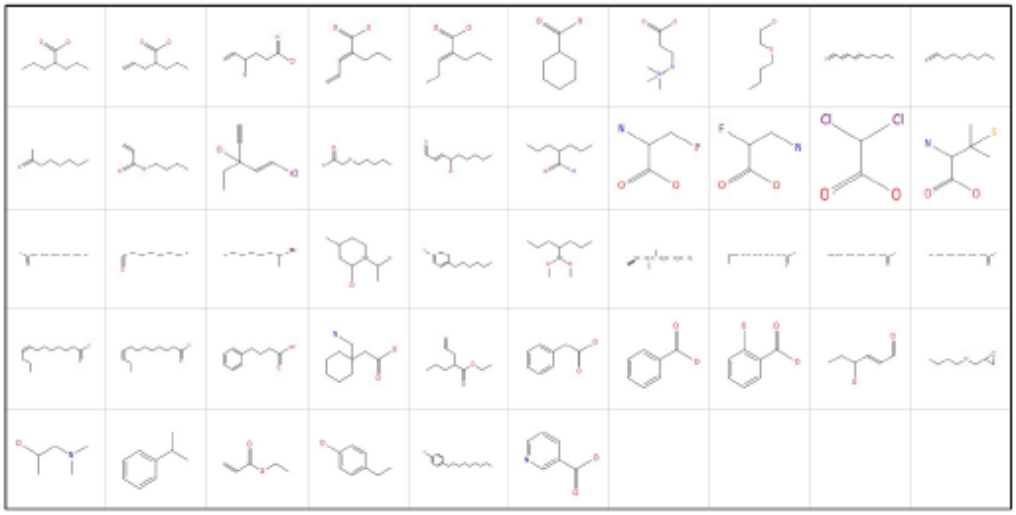
Representative 3D-Visualisations Using Principal Components 0, 1 and 2 (Physicochemical Descriptors) Showing Case Study 1 Compounds (Large Purple Spheres) lying within the Chemical Space of the Lhasa Limited Metabolism Dataset (Small Red Spheres).



Note that in this table the following compounds have the same values: 2-methylhexanoic acid/2- ethylpentanoic acid and 2-ethylheptanoic acid/2-propylhexanoic acid; therefore, their points overlap in the plot giving nine points and not the expected eleven.

The Lhasa Limited metabolism dataset contains 2739 unique parent compounds. However, only a small proportion of these are used to furnish the supporting nearest neighbours in this analysis. Although most of them appear many times supporting the scores for various biotransformations, there are only 46 unique nearest neighbours (UNN) in this set. These are shown below:

Case Study Compounds in the Unique Nearest Neighbours (UNN) Set:

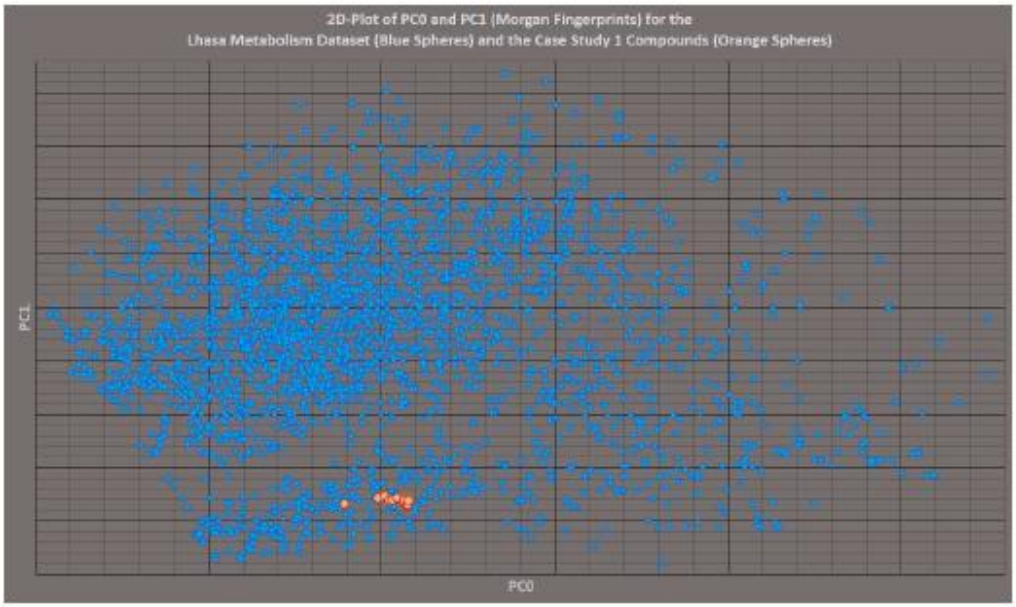


This set of compounds can be used to define a "localised" applicability domain in which the minimum and maximum values of principal components 0, 1 and 2 define a smaller area of chemical space than the whole dataset. All of the target and analogue compounds lie within this localised domain indicating that the nearest neighbours are physicochemically closer in nature to the query structures. This is to be expected as the choice of nearest neighbours within Meteor Nexus is constrained (by default) to those whose molecular weight is ±30% that of the query compound - molecular weight having some approximate covariance with other physicochemical descriptors such as LogP, total polar surface area, number of rotatable bonds and so on.

	PC0	PC1	PC2
Minimum value of Principal Component (UNN)	-6.07048	-6.6872	-5.10143
Maximum value of Principal Component (UNN)	7.234704	4.155276	3.932278
Minimum Value of Principal Component (Full Dataset)	-6.3509	-9.4763	-8.8195
Maximum Value of Principal Component (Full Dataset)	11.4787	5.5558	5.8887

Localised Applicability Domain (Purple Inner Box) Defined by the Set of 46 Unique Nearest Neighbours Compared to the Global Applicability Domain (Lilac Outer Box) Defined by the 2739 Compounds of the Lhasa Limited Metabolism Data Set. Case Study Compounds are Shown as Blue Spheres in Two-Dimensional Principal Component Space (P0 v. P1, P0 v. P2 and P1 v. P2)

The result of repeating the procedure of overlaying the data using the Morgan fingerprinting method is shown below:



Applicability Domain Declaration of the Mock Submission Target and Analogues Compounds Based on Physicochemical Principal Component Analysis and Morgan Fingerprints.

The analysis of physico-chemical and structural properties indicates that all compounds are covered in the domain of the Meteor Nexus knowledge base. This is summarised in the following table:

Compound Name	CAS Number	Applicability Domain	
		Phys. Chem (PCA)	Morgan Fingerprints
2-Ethylbutyric Acid	88-09-5	In Domain	In Domain
2-Propylheptanoic Acid	31080-39-4	In Domain	In Domain
2-Ethylheptanoic Acid	3274-29-1	In Domain	In Domain
2-Propylhexanoic Acid	3274-28-0	In Domain	In Domain
Valproic Acid	99-66-1	In Domain	In Domain
2-Ethylhexanoic Acid	149-57-5	In Domain	In Domain
2-Ethylpentanoic Acid	20225-24-5	In Domain	In Domain
2-Methylbutyric Acid	1730-91-2	In Domain	In Domain
2-Methylpentanoic Acid	97-61-0	In Domain	In Domain
2-Methylhexanoic Acid	4536-23-6	In Domain	In Domain
Pivalic Acid	75-98-9	In Domain	In Domain

Quality Definition of the Lhasa Limited Metabolism Dataset

Harvesting data regarding metabolic chemistry from the literature can be extremely challenging due the uncertainties inherent in the assignment of chemical structures and reaction pathways. Some of the elements contributing to difficulty of interpretation include: incomplete chromatographic and chemical characterisation, ambiguous or unknown regio- and stereochemistry, length and branching of reaction sequences, incomplete or ambiguous reaction sequence characterisation, obscure or unknown chemistry, and so on. Whilst automated methods in KNIME have been developed to check certain elements such as formatting integrity, much of our work has required expert peer review. A classification system has been developed which enables each paper to be assigned to a category

depending on the degree of difficulty in its interpretation. We have defined five categories as follows: "trivial: *easy and uncontentious*" (category 1), "straightforward: *slight ambiguity*" (category 2), "moderate: *need for some expert calls*" (category 3), "challenging: *need for many expert calls*" (category 4) and "advanced: *ambiguity remaining after expert calls*" (category 5). It is not practical to peer review every submission (and is not always needed) and so attention has been primarily focussed on categories 3-5, the target percentage of submissions subject to expert peer review increasing with the difficulty of interpretation. All submissions in category 5 have been subject to peer review by at least one principal or senior project team member. A random sample (1%) of submissions from each category was subjected to a further rigorous peer review and the errors and error rates noted. By extrapolation, the residual error rate across the database could be estimated. Here we report errors is the three categories assessed: *Experimental Protocol, References, Structures and Biotransformations*.

Residual Error Rate in the Lhasa Limited Metabolism Dataset version 2.2.0

Error Category	Typical Error Issue	Error Rate
Experimental Protocol	Species/strain assignment, dose levels, assay source, assay time, total recovery, amount of unchanged drug detected.	**11.90%**
References	Typographical errors in author list or paper title.	**1.28%**
Structures and Biotransformations	Incorrect parent compound or metabolite structure, misclassified biotransformation name, misclassified metabolite type (observed, presumed, expert call).	**2.02%**

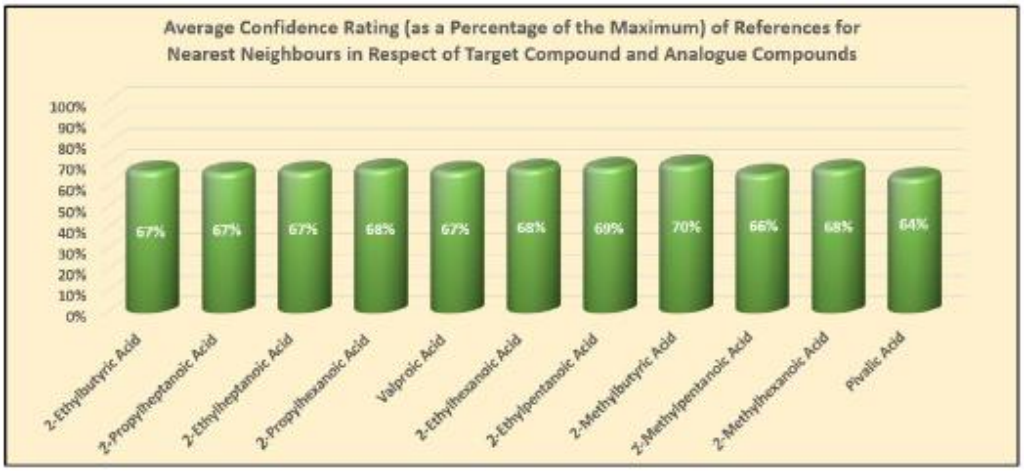
We had hoped for a residual error rate in all three categories of <5% and whilst the error rate in the *experimental protocol* category is higher than this, the information contained has no effect on the construction of the site-of-metabolism model from which nearest neighbours are chosen and upon which predicted biotransformation scores are predicated. This requires accuracy in the *structures and biotransformation* category and so we consider the residual error rate of 2% as acceptable.

Confidence Definition of References for Nearest Neighbour Examples Supporting Biotransformation Scores

This is expressed as an average across all references for each nearest neighbour example for each biotransformation for the target compound and the analogue compounds. Confidence ratings were calculated according to the method of Ponting *et al*[13]. The confidence rating is the summation of the *Chemical Characterisation Techniques Score* (0-4), the *Pathway Uncertainty Score* (0-2) and the *Metabolite Uncertainty Score* (0-2); the minimum combined score is 0 and the maximum 8. These are expressed as a percentage of the maximum. Individual rating definitions for each parameter are given in the table below:

Scheme of Assessment of Chemical Characterisation Confidence Ratings

Score / Parameter	0	1	2	3	4
Chemical Characterisation Techniques	Insufficient	Poor	Fair	Good	Excellent
Pathway Uncertainty	Sequence assignment has high level of presumed metabolites	Sequence assignment has moderate level of presumed metabolites	Sequence assignment has low level of presumed metabolites		
Metabolite Uncertainty	Structure assignment needs high level of expert call	Structure assignment needs moderate level of expert call	Structure assignment needs low level of expert call		



Individual Scores for Nearest Neighbour Examples in the Categories of: Chemical Characterisation Techniques, Pathway Uncertainty and Metabolite Uncertainty.

Comp_RefID	Techniques_verbal	Techniques_binned (0-4)	Pathway Uncertainty (0-2)	Metabolite Uncertainty (0-2)	SCORE (0-8)
Phenylbutyrate_094	excellent	4	2	2	8
Niacin_155	Poor	1	2	2	5
Phenylaceticacid_157	Insufficient	0	2	2	4
Cyclohexanecarboxylicacid_159	Poor	1	2	2	5
ButylGlycidylEther_233	Good	3	1	2	6
GSK977779_260	Good	3	2	0	5
L-menthol_415	Fair	2	2	2	6
2-Butoxyethanol_454	Poor	1	2	2	5
DCA_528	Good	3	2	2	7
R-2HMP_580	Fair	2	2	2	6
MET-88_603	Good	3	1	2	6
ButylAcrylate_859	Poor	1	1	2	4
EthylAcrylate_859	Poor	1	1	2	4
Trans-4-Hydroxy-2-hexenal_876	Good	3	2	2	7
4-Hydroxy-2-nonenal_881	Fair	2	2	2	6
ValproicAcid_920	Poor	1	1	2	4
VPA_1020	Excellent	4	2	2	8
VPA_1038	Fair	2	2	2	6
Butylphenol_1065	Poor	1	2	2	5
Ethylphenol_1065	Poor	1	2	2	5
Hexylphenol_1065	Poor	1	2	2	5
Nonylphenol_1065	Poor	1	2	2	5
2-trans-4-trans-Decadienal_1269	Good	3	2	2	7
Vigabatrin_1439	Good	3	2	2	7
10-DDNA_1492	Fair	2	2	2	6
11-DDNA_1492	Fair	2	1	2	5
8-DDNA_1492	Fair	2	2	2	6
9-DDNA_1492	Fair	2	2	2	6
LauricAcid_1492	Fair	2	2	2	6
2,4-dieneVPA_1499	Fair	2	2	2	6
Milacemide_1532	Fair	2	1	2	5
n-Butylacrylate_5008	Good	3	0	2	5
L-menthol_5037	Fair	2	1	2	5
Gabapentin_5062	Poor	1	2	2	5
2-Bromooctane_5115	Poor	1	2	2	5
2-Iodooctane_5115	Poor	1	2	2	5
2-Butoxyethanol_5116	Poor	1	2	2	5
Benzoicacid_5144	Fair	2	2	2	6
Trans-4-hydroxy-2-hexenal_5171	Fair	2	0	2	4
4-eneVPA_5181	Fair	2	1	2	5
Ethyl-4-eneVPA_5181	Fair	2	2	2	6
2-n-Propyl-4-Pentenoicacid_5210	Good	3	1	2	6
ValproicAcid_5212	Good	3	1	2	6

Fludalanine_5221	Fair	2	2	2	6
4-eneVPA_5232	Good	3	1	2	6
Penicillamine_5253	Fair	2	2	2	6
2-Propyl-2-pentenoicacid_5297	Good	3	0	2	5
PPDEA_5303	Fair	2	1	2	5
PPDIPA_5303	Fair	2	1	2	5
PPDMA_5303	Fair	2	1	2	5
Valpromide_5313	Insufficient	0	2	2	4
D-Penicillamine_5339	Poor	1	2	2	5
VPA_5386	Poor	1	2	2	5
DMAIP_5389	Fair	2	2	2	6
pAABA_5389	Fair	2	2	2	6
Doxifluridine_5409	Fair	2	0	2	4
alpha-Fluoro-beta-alanine_5409	Poor	1	2	2	5
Lauricacid_5551	Good	3	2	2	7
1-(2-chlorophenyl)propan-2-one_5614	Insufficient	0	2	2	4
1-(2-fluorophenyl)ethan-1-one_5614	Insufficient	0	2	2	4
1-(4-chlorophenyl)propan-2-one_5614	Insufficient	0	2	2	4
1-(o-tolyl)ethan-1-one_5614	Insufficient	0	2	2	4
1-(p-tolyl)propan-2-one_5614	Insufficient	0	2	2	4
1-phenylpropan-2-one_5614	Insufficient	0	2	2	4
acetophenone_5614	Insufficient	0	2	2	4
octan-2-one_5614	Insufficient	0	2	2	4
Octanal_5616	Insufficient	0	0	2	2
Ethchlorvynol_5624	Fair	2	1	2	5
SalicylicAcid_5638	Poor	1	2	2	5
SalicyluricAcid_5638	Poor	1	2	2	5
4-Nonylphenol-a_5786	Good	3	1	1	5
4-Nonylphenol-b_5786	Good	3	2	0	5
4-eneVPA_6173	Fair	2	1	2	5
4-eneVPA_6183	Good	3	1	2	6
2F-4-eneVPA_7066	Excellent	4	0	2	6

General Methods

Descriptions of Meteor and its reasoning (scoring) methods[14, 15,16,17] as well as discussions of some use cases[18,19] have been published previously.

Program Versions

Nexus v.2.2.0 (Build 52, Nov 2017)

Meteor Nexus v.3.1.0

Knowledge Base: Meteor KB 2018 1.0.0

Lhasa Limited Metabolism Data 2.2.0

Predicted Metabolism at First Generation Only

Process Constraints

Lhasa default except:

Max. Depth: 1

Score Threshold: 100

Inactive Biotransformations: 428

Report

Biotransformation Names, Biotransformation Numbers, and Absolute Scores for the target compound and each analogue compound were exported as a tsv file which was opened in Microsoft Excel for the purposes of creating a bar chart for ease of visualisation. For visual comparison of scores for analogue compound to target compound, the bar chart for the analogue compound was pasted onto a copy of that for the target compound.

Post Processing Filters

Score: >0 (tolerance ±0.0)

BIOTRANSFORMATION FINGERPRINT METHOD

We have defined the biotransformation fingerprint for a compound as a binary vector in which the occurrence of a given biotransformation is labelled as 1 and lack of occurrence of a given biotransformation is labelled as 0. In order to compare two molecules using the Tanimoto coefficient for binary data we need the fingerprints for the two molecules plus a third binary vector in which the occurrence of a given biotransformation in both molecules is recorded. In this analysis the occurrence of a biotransformation is defined as a biotransformation predicted with a score of >0.

Process Constraints

Lhasa default except:

Max. Depth: 1

Score Threshold: 100

Inactive Biotransformations: 428

Post Processing Filters

Score: >0 (tolerance ±0.0)

Example Calculation - Comparison of 2-Ethyl Butyric Acid to Valproic Acid

Biotransformation Name	A	B	AB
Glucuronidation of Carboxylic Acids	1	1	1
Conjugation of Alkyl Carboxylic Acids with Glycine	1	1	1
Conjugation of Carboxylic Acids with Glutamine	1	1	1
Conjugation of Carboxylic Acids with Taurine	0	0	0
Hydroxylation of Penultimate Alkyl Methylene	1	1	1
Hydroxylation of Penultimate Alkyl Methylene	1	1	1
Hydroxylation of Terminal Methyl	0	1	0
Hydroxylation of Terminal Methyl	0	1	0
Hydroxylation of Antepenultimate Alkyl Methylene	0	1	0
Hydroxylation of Antepenultimate Alkyl Methylene	0	1	0
PROTOTYPE - Glutathione Conjugation of Valproic Acid and Related Compounds	0	1	0
PROTOTYPE - Glutathione Conjugation of Valproic Acid and Related Compounds	0	1	0
Totals:	**5**	**11**	**5**

Where:

A = Biotransformations occurring for 2-ethylbutyric acid

B = Biotransformations occurring for valproic acid

AB = Biotransformations occurring for both 2-ethylbutyric acid and valproic acid

Similarity assessment using Tanimoto

An overall similarity assessment of the two biotransformation fingerprints can be generated by applying a Tanimoto equation that uses binary data:

$$\boldsymbol{Similarity = \frac{AB}{(A+B)-AB} = \frac{5}{(5+11)-5} = 0.46}$$

Coefficients were calculated between the target compound and all analogue compounds. These are conveniently visualised as a bar chart (comparing, for example, to the structural Tanimoto similarities). All analogues were also compared to each other and, together with the target compound, this is conveniently represented by a Tanimoto fingerprint matrix or heatmap.

Note on Assignment of Biotransformations for Both Methods:

For symmetric compounds like valproic acid, application of some biotransformations will generate only one metabolite (e.g. hydroxylation of terminal methyl) whilst those that are dissymmetric (like 2-propylheptanoic acid) will generate two metabolites for the same reaction, because the biotransformation operates twice - once on each different branch. For the purpose of generating the fingerprint, symmetric analogues like VPA are treated as if dissymmetric, that is, the biotransformation is scored twice. This is reasonable treatment - both carbon atoms in each propyl side-chain will react, it is just that the metabolites (ignoring any substrate/metabolite enantioselectivity) are identical. This makes for a "fairer" comparison. If both compounds to be compared are symmetric (e.g. 2-ethylbutyric and valproic) then one comparison is enough although for the fingerprint method we have scored twice to maintain parity across the analogue set. If both compounds are disymmetric then a four –fold comparison would be needed (1a:1b, 1a:2b, 2a:1b and 2b:2b) at least for the concordant metabolite method. As the target in this study is symmetrical this has not been needed, but this is a general feature of symmetry handling that needs to be considered as we automate the methods going forward.

References

1. Silva MFB, Aires CCP, Luis PBM, Ruiter JPN, Ijlst L, Duran M, Wanders RJA, and Tavares de Almeida I. Valproic Acid Metabolism and its Effects on Mitochondrial Fatty Acid Oxidation: A Review. *Journal of Inherited Metabolic Disease* (2008), 31, 205-216.
2. Ghodke-Puranik Y, Thorn CF, Lamba JK, Leeder JS, Song W, Birnbaum AK, Altman RB and Klein TE. Valproic Acid Pathway: Pharmacokinetics and Pharmacodynamics. *Pharmacogenetics and Genomics* (2013), 23, 236-241.
3. English JC, Deisinger PJ and Guest D. Metabolism of 2-Ethylhexanoic Acid Administered Orally or Dermally to the Female Fischer 344 Rat. *Xenobiotica* (1998), 28, 699-714.
4. Pennanen S, Auriola S, Manninen A and Komulainen H. Identification of the Main Metabolites of 2-Ethylhexanoic Acid in Rat Urine Using Gas Chromatography-Mass Spectrometry. *Journal of Chromatography B: Biomedical Sciences and Applications* (1991), 568, 125-134.
5. Pennanen S, Kojo A, Pasanen M, Liesivuori J, Juvonen RO and Komulainen H. CYP Enzymes Catalyze the Formation of a Terminal Olefin from 2-Ethylhexanoic Acid in Rat and Human Liver. *Human and Experimental Toxicology* (1996), 15, 435-442.
6. Walker V and Mills GA. Urine 4-Heptanone: a beta-Oxidation Product of 2-Ethylhexanoic Acid from Plasticisers. *Clinica Chimica Acta* (2001), 306, 51-61.
7. Hamdoune M, Duclos S, Mounie J, Santona L, Lhuguenot JC, Magdalou J and Goudonnet H. *In vitro* Glucuronidation of Peroxisomal Proliferators: 2-Ethylhexanoic Acid Enantiomers and Their Structural Analogs. *Toxicology and Applied Pharmacology* (1995), 131, 235-243.

8. Kanazu T and Yamaguchi T. Comparison of *in vitro* Carnitine and Glycine Conjugation with Branched-Side Chain and Cyclic Side Chain Carboxylic Acids in Rats. *Drug Metabolism and Disposition* (1997), 25, 149-153.
9. Kanazu T and Yamaguchi T. Substrate Specificity for Carnitine and Glycine Conjugation of Branched Side-Chain and Cyclic Side-Chain Carboxylic Acids in Various Experimental Animals. *Xenobiotica* (2009), 39, 335-344.
10. Xue L and Bajorath J. Molecular Descriptors for Effective Classification of Biologically Active Compounds Based on Principal Component Analysis Identified by a Genetic Algorithm. *Journal of Chemical Information and Computer Sciences* (2000), 40, 801-809.
11. Xue L and Bajorath J. Accurate Partitioning of Compounds Belonging to Diverse Activity Classes. *Journal of Chemical Information and Computer Sciences* (2002), 42, 757-764.
12. Rogers D and Hahn M. Extended-Connectivity Fingerprints. *Journal of Chemical Information and Modeling* (2010), 50, 742-754.
13. Ponting DJ, Murray E and Long A. Quantifying Confidence in the Reporting of Metabolic Biotransformations. *Drug Discovery Today* (2017), 22, 970-975.
14. Marchant CA, Briggs KA and Long A. *In silico* Tools for Sharing Data and Knowledge on Toxicity and Metabolism: Derek for Windows, Meteor, and Vitic. *Toxicology Mechanisms and Methods*, (2008), 18, 177-187.
15. https://www.lhasalimited.org/products/meteor-nexus.htm
16. Marchant CA, Rosser EM and Vessey JD. A k-Nearest Neighbours Approach Using Metabolism-Related Fingerprints to Improve *in silico* Metabolite Ranking. *Molecular Informatics*, (2017), 36.
17. Judson PN, Long A, Murray E and Patel M. Assessing Confidence in Predictions Using Veracity and Utility – A Case Study on the Prediction of Mammalian Metabolism by Meteor Nexus. *Molecular Informatics*, (2015), 34, 284-291.
18. Long A, Fielding K, McSweeney N, Payne MP and Smoraczewska E. Expert Systems: The Use of Expert Systems in Drug Design - Toxicity and Metabolism. (2012) in *Drug Design Strategies: Quantitative Approaches*, Livingstone DJ and Davis A (editors), Royal Society of Chemistry – Drug Discovery Series, pp. 279-344, ISBN: 978-1-84973-166-9.
19. Long A and Murray A. Prediction of Xenobiotic Metabolism. (2018) in *Applied Chemoinformatics: Achievements and Future Opportunities*, Gasteiger J and Engels T (editors), Wiley-VCH (2018), ISBN: 978-3-527-34201-3.
20. Hanser T, Barber C, Marchaland JF and Werner S. Applicability Domain: Towards a More Formal Definition. *SAR and QSAR in Environmental Research* (2016), 27, 893-909.

7. Models predicting the Molecular initiation events (MIEs)

We applied *in silico* QSAR models for predicting Molecular Initiating Events (MIEs) for all the 11 compounds of the read-across group.

For each MIE, four different models were applied. All four models were developed and validated using the same data retrieved from ToxCast, using four different modelling methods (https://www.epa.gov/chemicalresearch/toxicity-forecaster-toxcasttm-data):

1. Balanced Random Forest without feature selection (Chen *et al.*, 2004, http://statistics.berkeley.edu/sites/default/files/tech-reports/666.pdf);
2. Downsampling-based Random Forest (Zakharov *et al.*, 2014, JCIM, 54, 705-712) without feature selection;
3. Balanced random Forest + VSURF feature selection (Genuer *et al.*, 2015 R Journal, 7, 19-33);
4. Downsampling-based Random Forest + VSURF feature selection.

A total of nine endpoints addressing the up and down regulation of six transcription factors identified as MIE in the AOP network leading to hepatic steatosis (section 5.4) were modelled, for a total of 32 QSAR models.

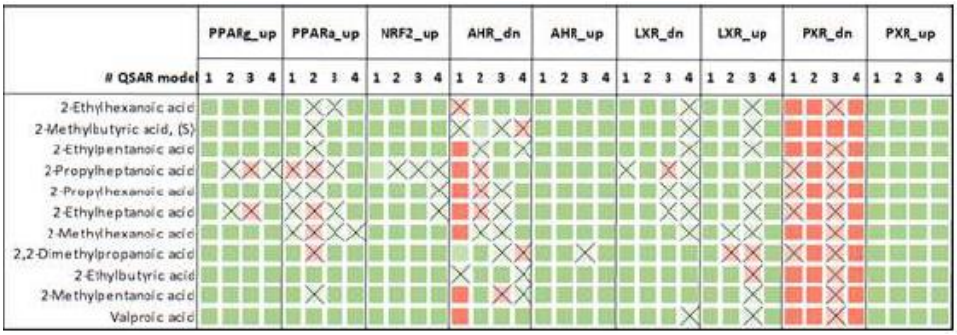
Further technical details on the four models are reported in QMRFs below.

Predictions are reported in Figure 1, in particular red squares flag positive responses, while green squares flag negative responses. Strikethrough squares flag predictions outside models' applicability domain (AD).

As shown, the majority of predictions returned from QSAR models are negatives. An exception is given by models for predicting Pregnane X receptor (PXR) down-regulation, that always returned positive outcomes.

Model 1 predicting Aryl Hydrocarbon Receptor (AHR) down-regulation gave positive predictions (in AD) for seven out of 11 chemicals. However, these positive predictions are never confirmed by other models. Indeed, three negative predictions (all in AD) are provided for Valproic acid with respect of a unique positive prediction. In some cases (e.g. 2-Propylhexanoic acid, 2-Ethylheptanoic acid and 2-Methylhxanoic acid) only two out of four predictions are within models' AD, being one positive and one negative, making difficult a final assessment of AHR down-regulation for these chemicals.

Figure 1. Prediction of MIES.



1) Balanced Random Forest without feature selection

QMRF identifier (JRC Inventory):To be entered by JRC

QMRF Title: QSARs for predicting up and down regulation of transcription factors as MIEs of hepatic steatosis (1)

Printing Date:26-giu-2018

1. QSAR identifier

1.1. QSAR identifier (title):

[1] BRF for predicting up regulation of pregnane X receptor (PXR)

[2] BRF RF for predicting down regulation of pregnane X receptor (PXR)

[3] BRF for predicting up regulation of liver X receptor (LXR)

[4] BRF for predicting down regulation of liver X receptor (LXR)

[5] BRF for predicting up regulation of Aryl hydrocarbon receptor (AhR)

[6] BRF for predicting down regulation of Aryl hydrocarbon receptor (AhR)

[7] BRF for predicting up regulation of Nuclear factor (erythroid-derived 2)-like 2 (Nrf2)

[8] BRF for predicting down regulation of Peroxisome proliferator-activated receptors alpha (PPARα)

[9] BRF for predicting down regulation of Peroxisome proliferator-activated receptors gamma (PPARγ)

1.2. Other related models:

1.3. Software coding the model:

RandomForest (R package) (v4.6-12).

KNIME (v3.4).

2. General information

2.1. Date of QMRF:

26 June 2018

2.2. QMRF author(s) and contact details:

Domenico Gadaleta; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; domenico.gadaleta@marionegri.it

2.3. Date of QMRF update(s):

2.4. QMRF update(s):

2.5. Model developer(s) and contact details:

[1] Domenico Gadaleta; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; domenico.gadaleta@marionegri.it

[2] Serena Manganelli; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; serena.manganelli@marionegri.it

[3] Cosimo Toma; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; cosimo.toma@marionegri.it

[4] Alessandra Roncaglioni; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; alessandra.roncaglioni@marionegri.it

[5] Emilio Benfenati; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; emilio.benfenati@marionegri.it

[6] Enrico Mombelli; Institut National de l'Environnement Industriel et des Risques (INERIS); enrico.mombelli@ineris.fr

2.6. Date of model development and/or publication:

2018

2.7. Reference(s) to main scientific papers and/or software package:

[1] Gadaleta, D., Manganelli, S., Roncaglioni, A., Toma, C., Benfenati, E., Mombelli, E. (2018). QSAR modelling of ToxCast assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. Journal of Chemical Information and Modelling, submitted manuscript.

[2] Judson, R.; Houck, K.; Martin, M.; Richard, A. M.; Knudsen, T. B.; Shah, I.; Little, S.; Wambaugh, J.; Woodrow Setzer, R.; Kothya, P.; Phuong, J.; Filer, D.; Smith, D.; Reif, D.; Rotroff, D.; Kleinstreuer, N.; Sipes, N.; Xia, M.; Huang, R.; Crofton, K.; Thomas, R. S., Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity on *In vitro* Assay Activity Across a Diverse Chemical and Assay Space. Toxicol. Sci. 2016, 152, 323-339.

[3] Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B., KNIME: The Konstanz Information Miner. In Data Analysis, Machine Learning and Applications, Preisach, C.; Burkhardt, H.; Schmidt-Thieme, L.; Decker, R., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp 319-326.

2.8. Availability of information about the model:

All the information about the model are reported in the reference publication (see 2.7).

2.9. Availability of another QMRF for exactly the same model:

3. Defining the endpoint - OECD Principle 1

3.1. Species: Human

3.2. Endpoint:

QMRF 6. Other QMRF 6. 6. Other

3.3. Comment on endpoint:

Up and/or down regulation of activity the following transcription factors:

[1] Pregnane X receptor (PXR) [GeneSymbol:NR1I2 | GeneID:8856 | Uniprot_SwissProt_Accession: O75469], up and down regulation

[2] Liver X recetor (LXR) [GeneSymbol:NR1H2 & NR1H3 | GeneID:7276 & 10062 | Uniprot_SwissProt_Accession: P55055 & Q13133], up and down regulation

[3] Aryl hydrocarbon receptor (AhR) [GeneSymbol: AHR | GeneID: 196 | Uniprot_SwissProt_Accession: P35869], up and down regulation

[4] Nuclear factor (erythroid-derived 2)-like 2 (Nrf2) [GeneSymbol: NFE2L2 | GeneID: 4780 | Uniprot_SwissProt_Accession: Q16236], up regulation

[5] Peroxisome proliferator-activated receptors alpha (PPARα) [GeneSymbol: PPARA | GeneID: 5465 | Uniprot_SwissProt_Accession: Q07869], up regulation

[6] Peroxisome proliferator-activated receptors gamma (PPARγ) [GeneSymbol: PPARA | GeneID: 5468 | Uniprot_SwissProt_Accession: P37231], up regulation

3.4. Endpoint units:

Adimensional

3.5. Dependent variable:

Nine endoints. Categorical (1 for positive, 0 for negative). pAC50 values greater than zero are actives (1), pAC50 values equal to zero are inactives (0). If at least one of the assays considered for each endpoint were active, the sample was flagged as active.

3.6. Experimental protocol:

Assays from ToxCast considered for each endpoint are the following:

[1] PXR up regulation: ATG_PXR_TRANS_up (AEID: 135), and ATG_PXRE_CIS_up (AEID: 103)

[2] PXR down regulation: ATG_PXR_TRANS_dn (AEID: 1475), and ATG_PXRE_CIS_dn (AEID: 1474)

[3] LXR up regulation: ATG_LXRa_TRANS_up (AEID: 125), ATG_LXRb_TRANS_up (AEID: 126), ATG_DR4_LXR_CIS_up (AEID: 70)

[4] LXR down regulation: ATG_LXRa_TRANS_dn (AEID: 1443), ATG_LXRb_TRANS_dn (AEID: 1444), ATG_DR4_LXR_CIS_dn (AEID: 1412)

[5] Ahr up regulation: ATG_Ahr_CIS_up (AEID: 63)

[6] Ahr down regulation: ATG_Ahr_CIS_dn (AEID: 1400)

[7] NrF2 up regulation: ATG_NRF2_ARE_CIS_up (AEID: 97)

[8] PPARα up regulation: ATG_PPARa_TRANS_up (AEID: 132)

[9] PPARγ up regulation: ATG_PPARg_TRANS_up (AEID: 134)

Attagene (ATG) assays are cell-based, multiplexed-redout assays that uses HepG2, a human liver cell line, with measurements taken at 24 hour after chemical dosing in 24-well plate.

These assays are designed to make measurements of mRNA induction, a form of inducible reporter, as detected with fluorescence intensity signals by Reverse transcription polymerase chain reaction (RT-PCR) and Capillary electrophoresis technology.

Changes to fluorescence intensity signals are indicative of inducible changes in transcription factor activity. This is quantified by the level of mRNA reporter sequence unique to:

The cis-acting reporter gene response element which are responsive of an endogenous human receptor subfamily (CIS assays);

The transfected trans-acting reporter gene and exogenous transcription factor GAL4, which are responsive of a given human receptor isoform (TRANS assays).

Further info on the assays: http://www.attagene.com/technology.php

3.7. Endpoint data quality and variability:

Experimental data used in this work were isolated from a collection of 24 *in vitro* HTS assays from the ToxCast program, executed by Attagene Inc. (RTP, NC), under contract to the U.S. EPA (Contract Number EP-W-07-049). During this program, several experiments evaluated the impact of more than 8,000 chemicals on the previously described TFs involved in the MIE of steatosis AOP.

For approximately half the chemicals tested during the ToxCast project, cytotoxicity was observed in the range of concentrations tested. Thus, a significant proportion of measured activities may represent a false positive response caused by assay interference process linked to a cytotoxicity-related 'burst' of activities (Judson *et al.*, 2006, Toxicol. Sci. 2016, 152, 323-339).

The presence of possible false negatives was also reported. The volatility of particular chemical categories (e.g., solvent chemicals) included in ToxCast or the low solubility may explain their general lack of significant effect.

4. Defining the algorithm - OECD Principle 2

4.1. Type of model:

Balanced Random Forest (BRF)

4.2. Explicit algorithm:

Balanced Random Forest (BRF) is a combination of under-sampling and the ensemble idea. This technique artificially alters the class distribution so that classes are represented equally in each tree. The randomForest R package (version 4.6-12) was used for the BRF approach. The *mtry* value was the one provided by default in R. The number of trees was selected in the range 25-501, based on the lowest prediction error returned in 10-fold internal cross validation. Table indicates the number of trees for each model:

PXR		LXR		AhR		NrF2	PPARγ	PPARα
up	dn	up	dn	up	dn	up	up	up
501	101	201	201	501	501	201	501	201

4.3. Descriptors in the model:

PXR_up: 1095

PXR_dn: 1127

LXR_up: 1134

LXR_dn: 1116

AhR_up: 1130

AhR_dn: 1126

NrF2_up: 1112

PPARg_up: 937

PPARa_up: 1126

4.4. Descriptor selection:

Automated.

4.5. Algorithm and descriptor generation:

Descriptors were pruned by constant and semi-constant vales (i.e. standard deviation < 0.01), then if a couple of descriptors was characterised by an absolute pair correlation greater than 90%, the descriptor with the highest pair correlation with all the other descriptors was removed. RF automatically identified descriptors most relevant for describing the endpoint.

4.6. Software name and version for descriptor generation:

Dragon v7.0.8

Calculation of several sets of molecular descriptors from molecular geometries (topological, geometrical, WHIM, 3D-MoRSE, molecular profiles, etc.)

Kode srl. Via Nino Pisano, 14 56122 Pisa (PI) - Italy, info@kode-solutions.net www.kodesolutions.net

https://chm.kode-solutions.net/products_dragon.php

4.7. Chemicals/Descriptors ratio:

Not relevant for Random Forests

5. Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model:

[1] Threshold on confidence of RF

[2] Threshold on similarity of nearest neighbors of the training set

5.2. Method used to assess the applicability domain:

ADs were optimised for each model by fine-tuning the two metrics with respect to predictivity:

[1] For the first AD criterion we estimated the percentage of trees within the RF yielding the same prediction (i.e. confidence). A confidence threshold (Tc) was implemented in KNIME and gradually incremented by 0.05, from 0.55 to 0.75. Chemicals with confidence lower than this threshold were considered as outside the model AD.

[2] The second AD criterion took account of the structural domain of the model. This was achieved by evaluating the degree of structural similarity of a given compound to those included within the TS. A distance matrix containing Euclidean distances for each pair of chemicals in the TS was calculated, then for each TS compound the mean distance from its first k neighbors was calculated. TS chemicals were then sorted on the basis of these distances and the value corresponding to a given percentile of the distribution of distances was used as a threshold (TD) beyond which chemicals were excluded from the AD. For the external validations, the same procedure was repeated calculating the prediction confidence and the distances of each VS chemical from their neighbors within the TS, then TD was used to identify chemicals outside of AD. For the present work, we adopted the Euclidean distance calculated on the scaled and centered descriptors used by the models as a similarity criterion; values assigned to k were 1 and 5; values assigned to TD were those corresponding to the 100th, the 97.5th, the 95th and the 90th percentiles of the TS distance distributions.

The best values for parameters were selected based on performance and coverage (i.e., percentage of predictions in the AD) in 10-fold-cross-validation. The final values are:

	PXR		LXR		AhR		NrF2	PPARγ	PPARα
	up	dn	up	dn	up	dn	up	up	up
T_D	0.65	0.65	0.60	0.70	0.65	0.60	0.65	0.60	0.65
T_C	100th	95th	95th	90th	100th	100th	95th	100th	90th
MN	1	5	5	1	1	1	1	1	5

5.3. Software name and version for applicability domain assessment:

KNIME (Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization); Prof. Dr. Michael Berthold, Michael.Berthold@uni-konstanz.de https://www.knime.com/

5.4. Limits of applicability:

The model is not applicable on inorganic chemicals and those including unusual elements (i.e., different from C, O, N, S, Cl, Br, F, I). Salts can be predicted only if stripped of the counterion and converted to the neutralised form.

6. Internal validation - OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Available information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

Not available

6.4. Data for the dependent variable for the training set:

All

6.5. Other information about the training set:

Datasets for each endpoint were randomly divided into a training set (TS, 80% of the original dataset) and a validation set (VS, 20% of the original dataset) comprising the same proportion of active and inactive chemicals as the original dataset. The table below reports the number of chemicals in TS and VS for each of the modelled endpoint.

TF	Inactives (full database)	Actives (full database)	% Actives	TS	VS
PXR_up	529	640	55	934	235
PXR_dn	1269	83	6	1079	273
LXR_up	1296	68	5	1089	275
LXR_dn	990	141	12	904	227
AhR_up	1148	162	12	1045	265
AhR_dn	1301	50	4	1079	272
NRF2_up	723	346	32	853	216
PPARγ_up	871	266	23	908	229
PPARα_u	1259	64	5	1057	266

6.6. Pre-processing of data before modelling:

Data were retrieved from the oldstyle_neg_log_ac50_Matrix_151020.csv file, downloaded from ftp://newftp.epa.gov/comptox/High_Throughput_Screening_Data/Summary_Files.

For classification purposes, for each assay chemicals with zero values for a given assay were considered as inactive, while chemicals with a continuous pAC50 value were considered active. Results from TRANS-assays were considered if specific isoforms of a TF listed among the AOPs for steatosis (e.g., PPARα, PPARγ), while CIS-assays or a combination of CIS- and TRANS-assays were used if TF isoforms were not specified. In the latter case, CIS- and TRANS- outputs were combined. A chemical was labelled as active if it was active in at least one assay and inactive if it was inactive in both types of assay.

Only chemicals exceeding 90% purity were retained, while chemicals associated to lower purity, other anomalies (e.g. withdrawn chemicals) or not yet analysed were not included.

The structures were checked by removing inorganic chemicals and mixtures, correcting inaccurate SMILES codes with the help of chemical databases, i.e. ChemSpider and ChemIDplus and neutralizing salts.

An in-house software was used to identify and remove duplicates. For a given set of duplicated structures, if their experimental activities were identical, then only one

compound was kept. If their experimental properties were different, both the chemicals were removed.

A Python script executing the MolVS standardiser (based on RDkit libraries) was written to obtain canonical tautomers. Canonical SMILES were coded using the istMolBase software based on CDK libraries.

A z-score (Eq. 1) that was assigned to each chemical-assay combination (Judson *et al.*, 2006, Toxicol. Sci. 2016, 152, 323-339):

$$Z(chemical\,,assay) = \frac{-\log AC_{50}(chemical\,,assay) - median[-\log AC_{50}(\text{chemical}\,,\text{cytotoxity})]}{global\ cytotoxicity\ MAD} \quad (1)$$

Conversely, chemicals associated with low z-scores are more likely to be false positives confounded by cytotoxicity. A z-score threshold of 3 was considered to select only chemicals that can be considered as specifically active.

6.7. Statistics for goodness-of-fit:

6.8. Robustness - Statistics obtained by leave-one-out cross-validation:

6.9. Robustness - Statistics obtained by leave-many-out cross-validation:

Method: 10-fold cross-validation. Results are reported above:

	PXR_up		PXR_dn		LXR_up		LXR_dn		AHR_up		AHR_dn		NRF2_up		PPARg_up		PPARa_up	
	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD
#	934	598	1079	616	1089	637	904	368	1045	507	1079	535	853	428	908	643	1057	546
P	512	341	66	27	54	30	112	37	129	65	40	18	276	130	212	123	50	23
N	422	257	1013	589	1035	607	792	331	916	442	1039	517	577	298	696	520	1007	523
ACC	0.73	0.80	0.80	0.91	0.83	0.93	0.64	0.78	0.69	0.80	0.55	0.56	0.67	0.76	0.76	0.83	0.87	0.95
SE	0.81	0.88	0.52	0.37	0.37	0.30	0.76	0.89	0.74	0.80	0.53	0.72	0.71	0.78	0.62	0.71	0.42	0.43
SP	0.64	0.70	0.82	0.94	0.86	0.96	0.62	0.77	0.69	0.80	0.55	0.56	0.66	0.76	0.81	0.86	0.90	0.97
MCC	0.46	0.59	0.20	0.23	0.14	0.26	0.25	0.43	0.30	0.44	0.03	0.10	0.34	0.50	0.40	0.52	0.21	0.40
BA	0.73	0.79	0.67	0.65	0.61	0.63	0.69	0.83	0.72	0.80	0.54	0.64	0.68	0.77	0.71	0.79	0.66	0.70
AUC	0.79	0.84	0.72	0.73	0.67	0.72	0.74	0.83	0.77	0.82	0.57	0.62	0.74	0.81	0.80	0.83	0.71	0.77
%	1.00	0.64	1.00	0.57	1.00	0.58	1.00	0.41	1.00	0.49	1.00	0.50	1.00	0.50	1.00	0.71	1.00	0.52

6.10. Robustness - Statistics obtained by Y-scrambling:

	PXR_up	PXR_dn	LXR_up	LXR_dn	AHR_up	AHR_dn	NRF2_up	PPARg_up	PPARa_up
MCC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

6.11. Robustness - Statistics obtained by bootstrap:

6.12. Robustness - Statistics obtained by other methods:

7. External validation - OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Available information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

No

7.4. Data for the dependent variable for the external validation set:

All

7.5. Other information about the external validation set:

See 6.5

7.6. Experimental design of test set:

See 6.5

7.7. Predictivity - Statistics obtained by external validation:

	PXR_up		PXR_dn		LXR_up		LXR_dn		AHR_up		AHR_dn		NRF2_up		PPARg_up		PPARa_up	
	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD
#	235	151	273	151	275	148	227	103	265	137	272	134	216	120	229	166	266	112
P	128	92	17	7	14	6	29	14	33	20	10	6	70	36	54	35	14	6
N	107	59	256	144	261	142	198	89	232	117	262	128	146	84	175	131	252	106
ACC	0.77	0.87	0.83	0.91	0.81	0.91	0.67	0.78	0.66	0.82	0.49	0.54	0.66	0.74	0.75	0.81	0.89	0.96
SE	0.88	0.93	0.41	0.29	0.36	0.33	0.69	0.64	0.61	0.70	0.90	1.00	0.64	0.75	0.76	0.83	0.71	0.50
SP	0.64	0.76	0.86	0.94	0.83	0.94	0.67	0.80	0.66	0.84	0.47	0.52	0.67	0.74	0.75	0.81	0.90	0.98
MCC	0.53	0.72	0.18	0.19	0.11	0.20	0.25	0.34	0.18	0.44	0.14	0.22	0.30	0.46	0.45	0.56	0.40	0.52
BA	0.76	0.85	0.63	0.62	0.59	0.63	0.68	0.72	0.63	0.77	0.69	0.76	0.66	0.74	0.75	0.82	0.81	0.74
AUC	0.85	0.87	0.74	0.89	0.63	0.54	0.71	0.82	0.73	0.83	0.75	0.65	0.72	0.81	0.83	0.88	0.77	0.73
%	1.00	0.64	1.00	0.55	1.00	0.54	1.00	0.45	1.00	0.52	1.00	0.49	1.00	0.56	1.00	0.72	1.00	0.42

7.8. Predictivity - Assessment of the external validation set:

MCC values were lower in external validation with respect of internal validation for endpoints with highly unbalanced datasets, as in the case of LXR_up and AhR_dn (MCC < 0.30) for chemicals in the AD. A possible explanation for this poor performance, can be found in the extreme degree of imbalance of some VS (i.e. less than 10% of active chemicals) that seriously undermines the reliability of statistical indicators.

Statistical analyses were done to identify critical MCC thresholds for reliably evaluating the performance of models on binary datasets with different degree of imbalance. These thresholds correspond to a reasonable minimum predictivity and were defined for each model by imposing a minimum percentage of correctly predicted positive and negative

chemicals of 75% (i.e. SE = SP=75%). Results demonstrated that, given the same percentage of correctly predicted active and inactive compounds, very unbalanced datasets are linked to lower MCC values. Table below shows which models overcome predictivity thresholds with respect of the degree of unbalance of datasets.

	PXR_up	PXR_dn	LXR_up	LXR_dn	AHR_up	AHR_dn	NRF2_up	PPARg_up	PPARa_up
#	598	616	637	368	507	535	428	643	546
P	341	27	30	37	65	18	130	123	23
N	257	589	607	331	442	517	298	520	523
MCC	0.59	0.23	0.26	0.43	0.44	0.1	0.5	0.52	0.4
MCC75	0.50	0.23	0.25	0.33	0.36	0.22	0.47	0.41	0.22
Valid?	Y	Y	Y	Y	Y	N	Y	Y	Y
#	151	151	148	103	137	134	120	166	112
P	92	7	6	14	20	6	36	35	6
N	59	144	142	89	117	128	84	131	106
MCC	0.72	0.19	0.2	0.34	0.44	0.22	0.46	0.56	0.52
MCC75	0.49	0.22	0.26	0.40	0.38	0.27	0.47	0.42	0.30
Valid?	Y	Y/N	N	N	Y	N	Y/N	Y	Y

7.9. Comments on the external validation of the model:

8. Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model:

Not provided

8.2. A priori or a posteriori mechanistic interpretation:

8.3. Other information about the mechanistic interpretation:

9. Miscellaneous information

9.1. Comments:

9.2. Bibliography:

[1] Gadaleta, D., Manganelli, S., Roncaglioni, A., Toma, C., Benfenati, E., Mombelli, E. (2018). QSAR modelling of ToxCast assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. Journal of Chemical Information and Modelling, *submitted manuscript*.

[2] Judson, R.; Houck, K.; Martin, M.; Richard, A. M.; Knudsen, T. B.; Shah, I.; Little, S.; Wambaugh, J.; Woodrow Setzer, R.; Kothya, P.; Phuong, J.; Filer, D.; Smith, D.; Reif, D.; Rotroff, D.; Kleinstreuer, N.; Sipes, N.; Xia, M.; Huang, R.; Crofton, K.; Thomas, R. S., Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity on *In vitro* Assay Activity Across a Diverse Chemical and Assay Space. Toxicol. Sci. 2016, 152, 323-339.

[3] Romanov, S., Medvedev, A., Gambarian, M., Poltoratskaya, N., Moeser, M., Medvedeva, L., & Makarov, S. (2008). Homogeneous reporter system enables quantitative functional assessment of multiple transcription factors. Nature Methods, 5(3), 253.

[4] Martin, M. T., Dix, D. J., Judson, R. S., Kavlock, R. J., Reif, D. M., Richard, A. M., & Gambarian, M. (2010). Impact of environmental chemicals on key transcription regulators

and correlation to toxicity endpoints within EPA's ToxCast program. Chemical research in toxicology, 23(3), 578-590.

[5] Liaw, A.; Wiener, M., Classification and Regression by RandomForest. R News 2002, 2, 18-22.

9.3.Supporting information:

Training set(s)Test set(s)Supporting information

2) Downsampling-based Random Forest without feature selection QMRF

QMRF identifier (JRC Inventory):To be entered by JRC

QMRF Title: QSARs for predicting up and down regulation of transcription factors as MIEs of hepatic steatosis (2)

Printing Date:26-giu-2018

1. QSAR identifier

1.1. QSAR identifier (title):

[1] BRF for predicting up regulation of pregnane X receptor (PXR)

[2] BRF RF for predicting down regulation of pregnane X receptor (PXR)

[3] BRF for predicting up regulation of liver X receptor (LXR)

[4] BRF for predicting down regulation of liver X receptor (LXR)

[5] BRF for predicting up regulation of Aryl hydrocarbon receptor (AhR)

[6] BRF for predicting down regulation of Aryl hydrocarbon receptor (AhR)

[7] BRF for predicting up regulation of Nuclear factor (erythroid-derived 2)-like 2 (Nrf2)

[8] BRF for predicting down regulation of Peroxisome proliferator-activated receptors alpha (PPARα)

[9] BRF for predicting down regulation of Peroxisome proliferator-activated receptors gamma (PPARγ)

1.2. Other related models:

1.3. Software coding the model:

KNIME (v3.4).

2. General information

2.1. Date of QMRF:

26 June 2018

2.2. QMRF author(s) and contact details:

Domenico Gadaleta; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; domenico.gadaleta@marionegri.it

2.3. Date of QMRF update(s):

2.4. QMRF update(s):

2.5. Model developer(s) and contact details:

[1] Domenico Gadaleta; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; domenico.gadaleta@marionegri.it

[2] Serena Manganelli; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; serena.manganelli@marionegri.it

[3] Cosimo Toma; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; cosimo.toma@marionegri.it

[4] Alessandra Roncaglioni; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; alessandra.roncaglioni@marionegri.it

[5] Emilio Benfenati; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; emilio.benfenati@marionegri.it

[6] Enrico Mombelli; Institut National de l'Environnement Industriel et des Risques (INERIS); enrico.mombelli@ineris.fr

2.6. Date of model development and/or publication:

2018

2.7. Reference(s) to main scientific papers and/or software package:

[1] Gadaleta, D., Manganelli, S., Roncaglioni, A., Toma, C., Benfenati, E., Mombelli, E. (2018). QSAR modelling of ToxCast assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. Journal of Chemical Information and Modelling, submitted manuscript.

[2] Judson, R.; Houck, K.; Martin, M.; Richard, A. M.; Knudsen, T. B.; Shah, I.; Little, S.; Wambaugh, J.; Woodrow Setzer, R.; Kothya, P.; Phuong, J.; Filer, D.; Smith, D.; Reif, D.; Rotroff, D.; Kleinstreuer, N.; Sipes, N.; Xia, M.; Huang, R.; Crofton, K.; Thomas, R. S., Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity on *In vitro* Assay Activity Across a Diverse Chemical and Assay Space. Toxicol. Sci. 2016, 152, 323-339.

[3] Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B., KNIME: The Konstanz Information Miner. In Data Analysis, Machine Learning and Applications, Preisach, C.; Burkhardt, H.; Schmidt-Thieme, L.; Decker, R., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp 319-326.

2.8. Availability of information about the model:

All the information about the model are reported in the reference publications (see 2.7).

2.9. Availability of another QMRF for exactly the same model:

3. Defining the endpoint - OECD Principle 1

3.1. Species: Human

3.2. Endpoint:

QMRF 6. Other QMRF 6. 6. Other

3.3. Comment on endpoint:

Up and/or down regulation of activity the following transcription factors:

[1] Pregnane X receptor (PXR) [GeneSymbol:NR1I2 | GeneID:8856 | Uniprot_SwissProt_Accession: O75469], up and down regulation

[2] Liver X recetor (LXR) [GeneSymbol:NR1H2 & NR1H3 | GeneID:7276 & 10062 | Uniprot_SwissProt_Accession: P55055 & Q13133], up and down regulation

[3] Aryl hydrocarbon receptor (AhR) [GeneSymbol: AHR | GeneID: 196 | Uniprot_SwissProt_Accession: P35869], up and down regulation

[4] Nuclear factor (erythroid-derived 2)-like 2 (Nrf2) [GeneSymbol: NFE2L2 | GeneID: 4780 | Uniprot_SwissProt_Accession: Q16236], up regulation

[5] Peroxisome proliferator-activated receptors alpha (PPARα) [GeneSymbol: PPARA | GeneID: 5465 | Uniprot_SwissProt_Accession: Q07869], up regulation

[6] Peroxisome proliferator-activated receptors gamma (PPARγ) [GeneSymbol: PPARA | GeneID: 5468 | Uniprot_SwissProt_Accession: P37231], up regulation

3.4. Endpoint units:

Adimensional

3.5. Dependent variable:

Nine endpoints. Categorical (1 for positive, 0 for negative). pAC50 values greater than zero are actives (1), pAC50 values equal to zero are inactives (0). If at least one of the assays considered for each endpoint were active, the sample was flagged as active.

3.6. Experimental protocol:

Assays from ToxCast considered for each endpoint are the following:

[1] PXR up regulation: ATG_PXR_TRANS_up (AEID: 135), and ATG_PXRE_CIS_up (AEID: 103)

[2] PXR down regulation: ATG_PXR_TRANS_dn (AEID: 1475), and ATG_PXRE_CIS_dn (AEID: 1474)

[3] LXR up regulation: ATG_LXRa_TRANS_up (AEID: 125), ATG_LXRb_TRANS_up (AEID: 126), ATG_DR4_LXR_CIS_up (AEID: 70)

[4] LXR down regulation: ATG_LXRa_TRANS_dn (AEID: 1443), ATG_LXRb_TRANS_dn (AEID: 1444), ATG_DR4_LXR_CIS_dn (AEID: 1412)

[5] Ahr up regulation: ATG_Ahr_CIS_up (AEID: 63)

[6] Ahr down regulation: ATG_Ahr_CIS_dn (AEID: 1400)

[7] NrF2 up regulation: ATG_NRF2_ARE_CIS_up (AEID: 97)

[8] PPARα up regulation: ATG_PPARa_TRANS_up (AEID: 132)

[9] PPARγ up regulation: ATG_PPARg_TRANS_up (AEID: 134)

Attagene (ATG) assays are cell-based, multiplexed-redout assays that uses HepG2, a human liver cell line, with measurements taken at 24 hour after chemical dosing in 24-well plate.

These assays are designed to make measurements of mRNA induction, a form of inducible reporter, as detected with fluorescence intensity signals by Reverse transcription polymerase chain reaction (RT-PCR) and Capillary electrophoresis technology.

Changes to fluorescence intensity signals are indicative of inducible changes in transcription factor activity. This is quantified by the level of mRNA reporter sequence unique to:

The cis-acting reporter gene response element which are responsive of an endogenous human receptor subfamily (CIS assays);

The transfected trans-acting reporter gene and exogenous transcription factor GAL4, which are responsive of a given human receptor isoform (TRANS assays).

Further info on the assays: http://www.attagene.com/technology.php

3.7. Endpoint data quality and variability:

Experimental data used in this work were isolated from a collection of 24 *in vitro* HTS assays from the ToxCast program, executed by Attagene Inc. (RTP, NC), under contract to the U.S. EPA (Contract Number EP-W-07-049). During this program, several experiments evaluated the impact of more than 8,000 chemicals on the previously described TFs involved in the MIE of steatosis AOP.

For approximately half the chemicals tested during the ToxCast project, cytotoxicity was observed in the range of concentrations tested. Thus, a significant proportion of measured activities may represent a false positive response caused by assay interference process linked to a cytotoxicity-related 'burst' of activities (Judson *et al.*, 2006, Toxicol. Sci. 2016, 152, 323-339).

The presence of possible false negatives was also reported. The volatility of particular chemical categories (e.g., solvent chemicals) included in ToxCast or the low solubility may explain their general lack of significant effect.

4. Defining the algorithm - OECD Principle 2

4.1. Type of model:

Random Forest (RF)

4.2. Explicit algorithm:

Random Forest was derived based on undersampling of the training set, i.e. random deletion of the most represented class (i.e. negative chemicals) until both classes were equal in number. This approach generated a dataset more suitable for treatment with classical machine learning methods. RF implemented in KNIME was used to derive the undersampling based model. The mtry value was the one provided by default in KNIME. The number of trees was selected in the range 25-251, based on the lowest prediction error returned in 10-fold internal cross validation. Table indicates the number of trees for each model:

PXR		LXR		AhR		NrF2	PPARγ	PPARα
up	dn	up	dn	up	dn	up	up	up
101	101	101	101	101	101	101	101	101

4.3. Descriptors in the model:

[1] PXR_up: 313

[2] PXR_dn: 302

[3] LXR_up: 310

[4] LXR_dn: 318

[5] AhR_up: 318

[6] AhR_dn: 270

[7] NrF2_up:320

[8] PPARg_up: 316

[9] PPARa_up: 282

4.4. Descriptor selection:

Automated.

4.5. Algorithm and descriptor generation:

Descriptors were pruned by constant and semi-constant vales (i.e. standard deviation < 0.01), then if a couple of descriptors was characterised by an absolute pair correlation greater than 90%, the descriptor with the highest pair correlation with all the other descriptors was removed. RF automatically identified descriptors most relevant for describing the endpoint.

4.6. Software name and version for descriptor generation:

Dragon v7.0.8

Calculation of several sets of molecular descriptors from molecular geometries (topological, geometrical, WHIM, 3D-MoRSE, molecular profiles, etc.)

Kode srl. Via Nino Pisano, 14 56122 Pisa (PI) - Italy, info@kode-solutions.net www.kodesolutions.net

https://chm.kode-solutions.net/products_dragon.php

4.7. Chemicals/Descriptors ratio:

Not relevant for Random Forests

5. Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model:

[1] Threshold on confidence of RF

[2] Threshold on similarity of nearest neighbors of the training set

5.2. Method used to assess the applicability domain:

ADs were optimised for each model by fine-tuning the two metrics with respect to predictivity:

[1] For the first AD criterion we estimated the percentage of trees within the RF yielding the same prediction (i.e. confidence). A confidence threshold (Tc) was implemented in KNIME and gradually incremented by 0.05, from 0.55 to 0.75. Chemicals with confidence lower than this threshold were considered as outside the model AD.

[2] The second AD criterion took account of the structural domain of the model. This was achieved by evaluating the degree of structural similarity of a given compound to those included within the TS. A distance matrix containing Euclidean distances for each pair of chemicals in the TS was calculated, then for each TS compound the mean distance from its first k neighbors was calculated. TS chemicals were then sorted on the basis of these distances and the value corresponding to a given percentile of the distribution of distances was used as a threshold (TD) beyond which chemicals were excluded from the AD. For the external validations, the same procedure was repeated calculating the prediction confidence and the distances of each VS chemical from their neighbors within the TS, then TD was used to identify chemicals outside of AD. For the present work, we adopted the Euclidean distance calculated on the scaled and centered descriptors used by the models as a similarity criterion; valuesassigned to k were 1 and 5; values assigned to TD were those corresponding to the 100th, the 97.5th, the 95th and the 90th percentiles of the TS distance distributions.

The best values for parameters were selected based on performance and coverage (i.e., percentage of predictions in the AD) in 10-fold-cross-validation. The final values are:

	PXR		LXR		AhR		NrF2	PPARγ	PPARα
	up	dn	up	dn	up	dn	up	up	up
T_D	0.60	0.65	0.60	0.70	0.65	0.55	0.65	0.60	0.65
T_C	97.5th	100th	90th	90th	100th	100th	97.5th	100th	90th
MN	5	1	1	1	1	1	1	1	1

5.3. Software name and version for applicability domain assessment:

KNIME (Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization); Prof. Dr. Michael Berthold, Michael.Berthold@uni-konstanz.de https://www.knime.com/

5.4. Limits of applicability:

The model is not applicable on inorganic chemicals and those including unusual elements (i.e., different from C, O, N, S, Cl, Br, F, I). Salts can be predicted only if stripped of the counterion and converted to the neutralised form.

6. Internal validation - OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Available information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

Not available

6.4. Data for the dependent variable for the training set:

All

6.5. Other information about the training set:

Datasets for each endpoint were randomly divided into a training set (TS, 80% of the original dataset) and a validation set (VS, 20% of the original dataset) comprising the same proportion of active and inactive chemicals as the original dataset. The table below reports the number of chemicals in TS and VS for each of the modelled endpoint.

TF	Inactives (full database)	Actives (full database)	% Actives	TS	TS_{us}	VS
PXR_up	529	640	55	934	934	235
PXR_dn	1269	83	6	1079	132	273
LXR_up	1296	68	5	1089	108	275
LXR_dn	990	141	12	904	224	227
AhR_up	1148	162	12	1045	258	265
AhR_dn	1301	50	4	1079	80	272
NRF2_up	723	346	32	853	552	216
PPARγ_up	871	266	23	908	424	229
PPARα_u	1259	64	5	1057	100	266

6.6. Pre-processing of data before modelling:

Data were retrieved from the oldstyle_neg_log_ac50_Matrix_151020.csv file, downloaded from ftp://newftp.epa.gov/comptox/High_Throughput_Screening_Data/Summary_Files.

For classification purposes, for each assay chemicals with zero values for a given assay were considered as inactive, while chemicals with a continuous pAC50 value were considered active. Results from TRANS-assays were considered if specific isoforms of a TF listed among the AOPs for steatosis (e.g., PPARα, PPARγ), while CIS-assays or a combination of CIS- and TRANS-assays were used if TF isoforms were not specified. In the latter case, CIS- and TRANS- outputs were combined. A chemical was labelled as active if it was active in at least one assay and inactive if it was inactive in both types of assay.

Only chemicals exceeding 90% purity were retained, while chemicals associated to lower purity, other anomalies (e.g. withdrawn chemicals) or not yet analysed were not included.

The structures were checked by removing inorganic chemicals and mixtures, correcting inaccurate SMILES codes with the help of chemical databases, i.e. ChemSpider and ChemIDplus and neutralizing salts.

An in-house software was used to identify and remove duplicates. For a given set of duplicated structures, if their experimental activities were identical, then only one compound was kept. If their experimental properties were different, both the chemicals were removed.

A Python script executing the MolVS standardiser (based on RDkit libraries) was written to obtain canonical tautomers. Canonical SMILES were coded using the istMolBase software based on CDK libraries.

A z-score (Eq. 1) that was assigned to each chemical-assay combination (Judson *et al.*, 2006, Toxicol. Sci. 2016, 152, 323-339):

$$Z(chemical, assay) = \frac{-\log AC_{50}(chemical, assay) - median[-\log AC_{50}(\text{chemical}, \text{cytotoxity})]}{global\ cytotoxicity\ MAD} \quad (1)$$

Conversely, chemicals associated with low z-scores are more likely to be false positives confounded by cytotoxicity. A z-score threshold of 3 was considered to select only chemicals that can be considered as specifically active.

6.7. Statistics for goodness-of-fit:

6.8. Robustness - Statistics obtained by leave-one-out cross-validation:

6.9. Robustness - Statistics obtained by leave-many-out cross-validation:

Method: 10-fold cross-validation. Results are reported above:

	PXR_up		PXR_dn		LXR_up		LXR_dn		AHR_up		AHR_dn		NRF2_up		PPARg_up		PPARa_up	
	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD
#	934	709	132	61	108	54	224	100	258	134	80	66	552	263	424	310	100	40
P	512	392	66	26	54	27	112	45	129	69	40	29	276	133	212	151	50	27
N	422	317	66	35	54	27	112	55	129	65	40	37	276	130	212	159	50	13
ACC	0.73	0.79	0.63	0.77	0.53	0.61	0.64	0.82	0.64	0.74	0.60	0.62	0.66	0.73	0.72	0.77	0.68	0.78
SE	0.80	0.86	0.64	0.81	0.52	0.63	0.69	0.87	0.63	0.77	0.55	0.59	0.68	0.77	0.74	0.82	0.66	0.70
SP	0.64	0.70	0.62	0.74	0.54	0.59	0.59	0.78	0.65	0.71	0.65	0.65	0.64	0.69	0.70	0.73	0.70	0.92
MCC	0.45	0.58	0.26	0.54	0.06	0.22	0.28	0.65	0.28	0.48	0.20	0.23	0.33	0.47	0.43	0.55	0.36	0.59
BA	0.72	0.78	0.63	0.78	0.53	0.61	0.64	0.82	0.64	0.74	0.60	0.62	0.66	0.73	0.72	0.78	0.68	0.81
AUC	0.79	0.82	0.71	0.81	0.56	0.62	0.74	0.83	0.72	0.81	0.62	0.61	0.72	0.78	0.79	0.83	0.72	0.80
%	1.00	0.76	1.00	0.46	1.00	0.50	1.00	0.45	1.00	0.52	1.00	0.83	1.00	0.48	1.00	0.73	1.00	0.40

6.10. Robustness - Statistics obtained by Y-scrambling:

	PXR_up	PXR_dn	LXR_up	LXR_dn	AHR_up	AHR_dn	NRF2_up	PPARg_up	PPARa_up
MCC	0.00	0.00	0.01	-0.01	0.00	0.00	0.00	0.00	0.01

6.11. Robustness - Statistics obtained by bootstrap:

6.12. Robustness - Statistics obtained by other methods:

7. External validation - OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Available information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

No

7.4. Data for the dependent variable for the external validation set:

All

7.5. Other information about the external validation set:

See 6.5

7.6. Experimental design of test set:

See 6.5

7.7. Predictivity - Statistics obtained by external validation:

	PXR_up		PXR_dn		LXR_up		LXR_dn		AHR_up		AHR_dn		NRF2_up		PPARg_up		PPARa_up	
	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD
#	235	171	273	135	275	123	227	101	265	131	272	208	216	108	229	181	266	85
P	128	99	17	10	14	6	29	16	33	19	10	9	70	31	54	44	14	5
N	107	72	256	125	261	117	198	85	232	112	262	199	146	77	175	137	252	80
ACC	0.74	0.82	0.64	0.73	0.58	0.62	0.63	0.69	0.65	0.79	0.53	0.53	0.63	0.71	0.72	0.76	0.67	0.87
SE	0.84	0.90	0.82	0.80	0.64	0.67	0.69	0.69	0.52	0.74	0.90	1.00	0.59	0.77	0.85	0.93	0.71	0.80
SP	0.62	0.71	0.63	0.72	0.57	0.62	0.62	0.69	0.67	0.79	0.52	0.51	0.66	0.69	0.68	0.71	0.67	0.88
MCC	0.47	0.63	0.22	0.29	0.10	0.12	0.21	0.29	0.13	0.42	0.16	0.21	0.23	0.42	0.45	0.55	0.18	0.43
BA	0.73	0.80	0.72	0.76	0.61	0.64	0.66	0.69	0.59	0.77	0.71	0.76	0.62	0.73	0.77	0.82	0.69	0.84
AUC	0.83	0.87	0.74	0.69	0.62	0.60	0.67	0.71	0.70	0.81	0.74	0.72	0.70	0.79	0.82	0.83	0.75	0.88
%	1.00	0.73	1.00	0.49	1.00	0.45	1.00	0.44	1.00	0.49	1.00	0.76	1.00	0.50	1.00	0.79	1.00	0.32

7.8. Predictivity - Assessment of the external validation set:

MCC values were lower in external validation with respect of internal validation for endpoints with highly unbalanced datasets, as in the case of LXR_up and AhR_dn (MCC < 0.30) for chemicals in the AD. A possible explanation for this poor performance can be found in the extreme degree of imbalance of some VS (i.e. less than 10% of active chemicals) that seriously undermines the reliability of statistical indicators.

Statistical analyses were done to identify critical MCC thresholds for reliably evaluating the performance of models on binary datasets with different degree of imbalance. These thresholds correspond to a reasonable minimum predictivity and were defined for each model by imposing a minimum percentage of correctly predicted positive and negative

chemicals of 75% (i.e. SE = SP=75%). Results demonstrated that, given the same percentage of correctly predicted active and inactive compounds, very unbalanced datasets are linked to lower MCC values. Table below shows which models overcome predictivity thresholds with respect of the degree of unbalance of datasets.

	PXR_up	PXR_dn	LXR_up	LXR_dn	AHR_up	AHR_dn	NRF2_up	PPARg_up	PPARa_up
#	709	61	54	100	134	66	263	310	40
P	392	26	27	45	69	29	133	151	27
N	317	35	27	55	65	37	130	159	13
MCC	0.58	0.54	0.22	0.65	0.48	0.23	0.47	0.55	0.59
MCC75	0.50	0.51	0.48	0.50	0.51	0.51	0.51	0.50	0.48
Valid?	Y	Y	N	Y	Y/N	N	Y/N	Y	Y
#	171	135	123	101	131	208	108	181	85
P	99	10	6	16	19	9	31	44	5
N	72	125	117	85	112	199	77	137	80
MCC	0.63	0.29	0.12	0.29	0.42	0.21	0.42	0.55	0.43
MCC75	0.49	0.32	0.28	0.39	0.37	0.24	0.46	0.45	0.29
Valid?	Y	Y/N	N	N	Y	Y/N	N	Y	Y

7.9. Comments on the external validation of the model:

8. Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model:

Not provided

8.2. A priori or a posteriori mechanistic interpretation:

8.3. Other information about the mechanistic interpretation:

9. Miscellaneous information

9.1. Comments:

9.2. Bibliography:

[1] Gadaleta, D., Manganelli, S., Roncaglioni, A., Toma, C., Benfenati, E., Mombelli, E. (2018). QSAR modelling of ToxCast assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. Journal of Chemical Information and Modelling, *submitted manuscript*.

[2] Judson, R.; Houck, K.; Martin, M.; Richard, A. M.; Knudsen, T. B.; Shah, I.; Little, S.; Wambaugh, J.; Woodrow Setzer, R.; Kothya, P.; Phuong, J.; Filer, D.; Smith, D.; Reif, D.; Rotroff, D.; Kleinstreuer, N.; Sipes, N.; Xia, M.; Huang, R.; Crofton, K.; Thomas, R. S., Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity on *In vitro* Assay Activity Across a Diverse Chemical and Assay Space. Toxicol. Sci. 2016, 152, 323-339.

[3] Romanov, S., Medvedev, A., Gambarian, M., Poltoratskaya, N., Moeser, M., Medvedeva, L., & Makarov, S. (2008). Homogeneous reporter system enables quantitative functional assessment of multiple transcription factors. Nature Methods, 5(3), 253.

[4] Martin, M. T., Dix, D. J., Judson, R. S., Kavlock, R. J., Reif, D. M., Richard, A. M., & Gambarian, M. (2010). Impact of environmental chemicals on key transcription regulators

and correlation to toxicity end points within EPA's ToxCast program. Chemical research in toxicology, 23(3), 578-590.

9.3. Supporting information:

Training set(s)Test set(s)Supporting information

3) Balanced Random Forest with VSURF-based feature selection

QMRF identifier (JRC Inventory):To be entered by JRC

QMRF Title: QSARs for predicting up and down regulation of transcription factors as MIEs of hepatic steatosis (3)

Printing Date:26-giu-2018

1. QSAR identifier

1.1. QSAR identifier (title):

[1] BRF for predicting up regulation of pregnane X receptor (PXR)

[2] BRF RF for predicting down regulation of pregnane X receptor (PXR)

[3] BRF for predicting up regulation of liver X receptor (LXR)

[4] BRF for predicting down regulation of liver X receptor (LXR)

[5] BRF for predicting up regulation of Aryl hydrocarbon receptor (AhR)

[6] BRF for predicting down regulation of Aryl hydrocarbon receptor (AhR)

[7] BRF for predicting up regulation of Nuclear factor (erythroid-derived 2)-like 2 (Nrf2)

[8] BRF for predicting down regulation of Peroxisome proliferator-activated receptors alpha (PPARα)

[9] BRF for predicting down regulation of Peroxisome proliferator-activated receptors gamma (PPARγ)

1.2. Other related models:

1.3. Software coding the model:

RandomForest (R package) (v4.6-12).

KNIME (v3.4).

2. General information

2.1. Date of QMRF:

26 June 2018

2.2. QMRF author(s) and contact details:

Domenico Gadaleta; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; domenico.gadaleta@marionegri.it

2.3. Date of QMRF update(s):

2.4. QMRF update(s):

2.5. Model developer(s) and contact details:

[1] Domenico Gadaleta; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; domenico.gadaleta@marionegri.it

[2] Serena Manganelli; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; serena.manganelli@marionegri.it

[3] Cosimo Toma; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; cosimo.toma@marionegri.it

[4] Alessandra Roncaglioni; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; alessandra.roncaglioni@marionegri.it

[5] Emilio Benfenati; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; emilio.benfenati@marionegri.it

[6] Enrico Mombelli; Institut National de l'Environnement Industriel et des Risques (INERIS); enrico.mombelli@ineris.fr

2.6. Date of model development and/or publication:

2018

2.7. Reference(s) to main scientific papers and/or software package:

[1] Gadaleta, D., Manganelli, S., Roncaglioni, A., Toma, C., Benfenati, E., Mombelli, E. (2018). QSAR modelling of ToxCast assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. Journal of Chemical Information and Modelling, submitted manuscript.

[2] Judson, R.; Houck, K.; Martin, M.; Richard, A. M.; Knudsen, T. B.; Shah, I.; Little, S.; Wambaugh, J.; Woodrow Setzer, R.; Kothya, P.; Phuong, J.; Filer, D.; Smith, D.; Reif, D.; Rotroff, D.; Kleinstreuer, N.; Sipes, N.; Xia, M.; Huang, R.; Crofton, K.; Thomas, R. S., Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity on *In vitro* Assay Activity Across a Diverse Chemical and Assay Space. Toxicol. Sci. 2016, 152, 323-339.

[3] Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B., KNIME: The Konstanz Information Miner. In Data Analysis, Machine Learning and Applications, Preisach, C.; Burkhardt, H.; Schmidt-Thieme, L.; Decker, R., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp 319-326.

[4] Genuer, R.; Poggi, J. M.; Tuleau-Malot, C., VSURF: An R Package for Variable Selection Using Random Forests. The R Journal 2015, 7, 19-33.

2.8. Availability of information about the model:

All the information about the model are reported in the reference publications (see 2.7).

2.9. Availability of another QMRF for exactly the same model:

3. Defining the endpoint - OECD Principle 1

3.1. Species: Human

3.2. Endpoint:

QMRF 6. Other QMRF 6. 6. Other

3.3. Comment on endpoint:

Up and/or down regulation of activity the following transcription factors:

[1] Pregnane X receptor (PXR) [GeneSymbol:NR1I2 | GeneID:8856 | Uniprot_SwissProt_Accession: O75469], up and down regulation

[2] Liver X recetor (LXR) [GeneSymbol:NR1H2 & NR1H3 | GeneID:7276 & 10062 | Uniprot_SwissProt_Accession: P55055 & Q13133], up and down regulation

[3] Aryl hydrocarbon receptor (AhR) [GeneSymbol: AHR | GeneID: 196 | Uniprot_SwissProt_Accession: P35869], up and down regulation

[4] Nuclear factor (erythroid-derived 2)-like 2 (Nrf2) [GeneSymbol: NFE2L2 | GeneID: 4780 | Uniprot_SwissProt_Accession: Q16236], up regulation

[5] Peroxisome proliferator-activated receptors alpha (PPARα) [GeneSymbol: PPARA | GeneID: 5465 | Uniprot_SwissProt_Accession: Q07869], up regulation

[6] Peroxisome proliferator-activated receptors gamma (PPARγ) [GeneSymbol: PPARA | GeneID: 5468 | Uniprot_SwissProt_Accession: P37231], up regulation

3.4. Endpoint units:

Adimensional

3.5. Dependent variable:

Nine endoints. Categorical (1 for positive, 0 for negative). pAC50 values greater than zero are actives (1), pAC50 values equal to zero are inactives (0). If at least one of the assays considered for each endpoint were active, the sample was flagged as active.

3.6. Experimental protocol:

Assays from ToxCast considered for each endpoint are the following:

[1] PXR up regulation: ATG_PXR_TRANS_up (AEID: 135), and ATG_PXRE_CIS_up (AEID: 103)

[2] PXR down regulation: ATG_PXR_TRANS_dn (AEID: 1475), and ATG_PXRE_CIS_dn (AEID: 1474)

[3] LXR up regulation: ATG_LXRa_TRANS_up (AEID: 125), ATG_LXRb_TRANS_up (AEID: 126), ATG_DR4_LXR_CIS_up (AEID: 70)

[4] LXR down regulation: ATG_LXRa_TRANS_dn (AEID: 1443), ATG_LXRb_TRANS_dn (AEID: 1444), ATG_DR4_LXR_CIS_dn (AEID: 1412)

[5] Ahr up regulation: ATG_Ahr_CIS_up (AEID: 63)

[6] Ahr down regulation: ATG_Ahr_CIS_dn (AEID: 1400)

[7] NrF2 up regulation: ATG_NRF2_ARE_CIS_up (AEID: 97)

[8] PPARα up regulation: ATG_PPARa_TRANS_up (AEID: 132)

[9] PPARγ up regulation: ATG_PPARg_TRANS_up (AEID: 134)

Attagene (ATG) assays are cell-based, multiplexed-redout assays that uses HepG2, a human liver cell line, with measurements taken at 24 hour after chemical dosing in 24-well plate.

These assays are designed to make measurements of mRNA induction, a form of inducible reporter, as detected with fluorescence intensity signals by Reverse transcription polymerase chain reaction (RT-PCR) and Capillary electrophoresis technology.

Changes to fluorescence intensity signals are indicative of inducible changes in transcription factor activity. This is quantified by the level of mRNA reporter sequence unique to:

The cis-acting reporter gene response element which are responsive of an endogenous human receptor subfamily (CIS assays);

The transfected trans-acting reporter gene and exogenous transcription factor GAL4, which are responsive of a given human receptor isoform (TRANS assays).

Further info on the assays: http://www.attagene.com/technology.php

3.7. Endpoint data quality and variability:

Experimental data used in this work were isolated from a collection of 24 *in vitro* HTS assays from the ToxCast program, executed by Attagene Inc. (RTP, NC), under contract to the U.S. EPA (Contract Number EP-W-07-049). During this program, several experiments evaluated the impact of more than 8,000 chemicals on the previously described TFs involved in the MIE of steatosis AOP.

For approximately half the chemicals tested during the ToxCast project, cytotoxicity was observed in the range of concentrations tested. Thus, a significant proportion of measured activities may represent a false positive response caused by assay interference process linked to a cytotoxicity-related 'burst' of activities (Judson *et al.*, 2006, Toxicol. Sci. 2016, 152, 323-339).

The presence of possible false negatives was also reported. The volatility of particular chemical categories (e.g., solvent chemicals) included in ToxCast or the low solubility may explain their general lack of significant effect.

4. Defining the algorithm - OECD Principle 2

4.1. Type of model:

Balanced Random Forest (BRF)

4.2. Explicit algorithm:

Balanced Random Forest (BRF) is a combination of under-sampling and the ensemble idea. This technique artificially alters the class distribution so that classes are represented equally in each tree.

The randomForest R package (version 4.6-12) was used for the BRF approach. The mtry value was the one provided by default in KNIME. The number of trees was selected according to indications given by VSURF. Indeed, the package enables the identification of best descriptor subsets as a function of the number of trees. Table indicates the number of trees for each model:

PXR		LXR		AhR		NrF2	PPARγ	PPARα
up	dn	up	dn	up	dn	up	up	up
501	101	201	201	501	501	201	501	201

4.3. Descriptors in the model:

[1] PXR_up: ChiA_Dz.Z., Eta_alpha_A, GGI3, MAXDN, MLOGP2, P_VSA_ppp_D, P_VSA_ppp_L, SdssC, SpMax_B.m., SpMax_B.s., SpMax4_Bh.m., SpMin2_Bh.m., SpPosA_B.v., Wap, X5v

[2] PXR_dn: MLOGP, VE1_B.s., SpMaxA_B.s., X3v, P_VSA_ppp_L, IDM, MPC04, O%, SpMax5_Bh.m., P_VSA_v_3, ATSC7m, ATS8m, GATS2i, J_Dz.e., P_VSA_s_4, GGI7, ATSC2s, ATSC5e

[3] LXR_up: IC3, GATS8v, MAXDP, X2A, ATS4s, MLOGP2, SM15_EA.ri.

[4] LXR_dn: SpMin2_Bh.e., MLOGP, P_VSA_i_2, CATS2D_02_LL, SIC0, D.Dtr06, VE1_B.s., CATS2D_04_LL

[5] AhR_up: ARR, GATS1i, Eta_beta_A, SpMAD_B.v., GATS2e, MATS1p, SpPosA_B.v., AVS_B.e., SIC4, GATS2i, Eig13_AEA.dm., rGes, DLS_05

[6] AhR_dn: ATSC1s, VE1sign_B.m., SpMin3_Bh.v., IDDE, J_D.Dt

[7] NrF2_up: MLOGP, P_VSA_ppp_L, P_VSA_e_2, SpMin2_Bh.e., P_VSA_i_2, F07.C.C., NaasC, IDDE, CATS2D_06_LL, CATS2D_04_LL, IC3, ATSC2s, P_VSA_ppp_cyc

[8] PPARg_up: SpMAD_B.m., rGes, P_VSA_v_3, P_VSA_LogP_4, O%, O-057, nCconj, nCb., [1] MLOGP2, MLOGP, Mi, GATS8s, GATS1p, GATS1m, F06.C.O., D.Dtr06, CATS2D_07_AL, CATS2D_03_LL, C%, C-026, B07.C.O., ATSC7m

[9] PPARa_up: TI2_L, PW4, CATS2D_07_NL

4.4. Descriptor selection:

Automated. Use of VSURF (https://cran.r-project.org/web/packages/VSURF/VSURF.pdf) R package.

4.5. Algorithm and descriptor generation:

Descriptors were pruned by constant and semi-constant vales (i.e. standard deviation < 0.01), then if a couple of descriptors was characterised by an absolute pair correlation greater than 90%, the descriptor with the highest pair correlation with all the other descriptors was removed.

Optimal subsets of descriptors for modelling were obtained with the R package VSURF. The algorithm consists in a three step variable selection based on the logic underpinning the random forest (RF) algorithm (i.e. permutation importance and out-of-bag error). The first step eliminates irrelevant descriptors according to the permutation-based RF score of importance and a user-defined threshold. The second step finds important descriptors closely related to the response variable (interpretation step) and the third step (prediction step) identifies a sufficient parsimonious set of important descriptors leading to a good prediction of the response variables. The VSURF selection procedure was carried out as a function of a number of trees ranging from 25 to 251, then the pool of descriptors returning the lowest internal error was retained.

4.6. Software name and version for descriptor generation:

Dragon v7.0.8

Calculation of several sets of molecular descriptors from molecular geometries (topological, geometrical, WHIM, 3D-MoRSE, molecular profiles, etc.)

Kode srl. Via Nino Pisano, 14 56122 Pisa (PI) - Italy, info@kode-solutions.net www.kodesolutions.net

https://chm.kode-solutions.net/products_dragon.php

4.7. Chemicals/Descriptors ratio:

Not relevant for Random Forests

5. Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model:

[1] Threshold on confidence of RF

[2] Threshold on similarity of nearest neighbors of the training set

5.2. Method used to assess the applicability domain:

ADs were optimised for each model by fine-tuning the two metrics with respect to predictivity:

[1] For the first AD criterion we estimated the percentage of trees within the RF yielding the same prediction (i.e. confidence). A confidence threshold (Tc) was implemented in KNIME and gradually incremented by 0.05, from 0.55 to 0.75. Chemicals with confidence lower than this threshold were considered as outside the model AD.

[2] The second AD criterion took account of the structural domain of the model. This was achieved by evaluating the degree of structural similarity of a given compound to those included within the TS. A distance matrix containing Euclidean distances for each pair of chemicals in the TS was calculated, then for each TS compound the mean distance from its first k neighbors was calculated. TS chemicals were then sorted on the basis of these distances and the value corresponding to a given percentile of the distribution of distances was used as a threshold (TD) beyond which chemicals were excluded from the AD. For the external validations, the same procedure was repeated calculating the prediction confidence and the distances of each VS chemical from their neighbors within the TS, then TD was used to identify chemicals outside of AD. For the present work, we adopted the Euclidean

distance calculated on the scaled and centered descriptors used by the models as a similarity criterion; values assigned to k were 1 and 5; values assigned to TD were those corresponding to the 100th, the 97.5th, the 95th and the 90th percentiles of the TS distance distributions.

The best values for parameters were selected based on performance and coverage (i.e., percentage of predictions in the AD) in 10-fold-cross-validation. The final values are:

	PXR		LXR		AhR		NrF2	PPARγ	PPARα
	up	dn	up	dn	up	dn	up	up	up
T_D	0.65	0.70	0.60	0.60	0.70	0.65	0.65	0.60	0.75
T_C	90th	90th	95th	95th	90th	95th	95th	100th	100th
MN	1	1	5	1	1	5	1	1	1

5.3. Software name and version for applicability domain assessment:

KNIME (Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization); Prof. Dr. Michael Berthold, Michael.Berthold@uni-konstanz.de

https://www.knime.com/

5.4. Limits of applicability:

The model is not applicable on inorganic chemicals and those including unusual elements (i.e., different from C, O, N, S, Cl, Br, F, I). Salts can be predicted only if stripped of the counterion and converted to the neutralised form.

6. Internal validation - OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Available information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

Not available

6.4. Data for the dependent variable for the training set:

All

6.5. Other information about the training set:

Datasets for each endpoint were randomly divided into a training set (TS, 80% of the original dataset) and a validation set (VS, 20% of the original dataset) comprising the same proportion of active and inactive chemicals as the original dataset. The table below reports the number of chemicals in TS and VS for each of the modelled endpoint.

TF	Inactives (full database)	Actives (full database)	% Actives	TS	VS
PXR_up	529	640	55	934	235
PXR_dn	1269	83	6	1079	273
LXR_up	1296	68	5	1089	275
LXR_dn	990	141	12	904	227
AhR_up	1148	162	12	1045	265
AhR_dn	1301	50	4	1079	272
NRF2_up	723	346	32	853	216
PPARγ_up	871	266	23	908	229
PPARα_u	1259	64	5	1057	266

6.6. Pre-processing of data before modelling:

Data were retrieved from the oldstyle_neg_log_ac50_Matrix_151020.csv file, downloaded from ftp://newftp.epa.gov/comptox/High_Throughput_Screening_Data/Summary_Files.

For classification purposes, for each assay chemicals with zero values for a given assay were considered as inactive, while chemicals with a continuous pAC50 value were considered active. Results from TRANS-assays were considered if specific isoforms of a TF listed among the AOPs for steatosis (e.g., PPARα, PPARγ), while CIS-assays or a combination of CIS- and TRANS-assays were used if TF isoforms were not specified. In the latter case, CIS- and TRANS- outputs were combined. A chemical was labelled as active if it was active in at least one assay and inactive if it was inactive in both types of assay.

Only chemicals exceeding 90% purity were retained, while chemicals associated to lower purity, other anomalies (e.g. withdrawn chemicals) or not yet analysed were not included.

The structures were checked by removing inorganic chemicals and mixtures, correcting inaccurate SMILES codes with the help of chemical databases, i.e. ChemSpider and ChemIDplus and neutralizing salts.

An in-house software was used to identify and remove duplicates. For a given set of duplicated structures, if their experimental activities were identical, then only one compound was kept. If their experimental properties were different, both the chemicals were removed.

A Python script executing the MolVS standardiser (based on RDkit libraries) was written to obtain canonical tautomers. Canonical SMILES were coded using the istMolBase software based on CDK libraries.

A z-score (Eq. 1) that was assigned to each chemical-assay combination (Judson *et al.*, 2006, Toxicol. Sci. 2016, 152, 323-339):

$$Z(chemical, assay) = \frac{-\log AC_{50}(chemical, assay) - median[-\log AC_{50}(\text{chemical}, \text{cytotoxity})]}{global\ cytotoxicity\ MAD} \quad (1)$$

Conversely, chemicals associated with low z-scores are more likely to be false positives confounded by cytotoxicity. A z-score threshold of 3 was considered to select only chemicals that can be considered as specifically active.

6.7. Statistics for goodness-of-fit:

6.8. Robustness - Statistics obtained by leave-one-out cross-validation:

6.9. Robustness - Statistics obtained by leave-many-out cross-validation:

Method: 10-fold cross-validation. Results are reported above:

	PXR_up		PXR_dn		LXR_up		LXR_dn		AHR_up		AHR_dn		NRF2_up		PPARg_up		PPARa_up	
	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD
#	934	581	1079	482	1089	683	904	657	1045	463	1079	491	853	411	908	699	1057	608
P	512	322	66	26	54	36	112	70	129	51	40	16	276	110	212	151	50	26
N	422	259	1013	456	1035	647	792	587	916	412	1039	475	577	301	696	548	1007	582
ACC	0.74	0.81	0.63	0.77	0.79	0.87	0.61	0.67	0.77	0.89	0.57	0.59	0.70	0.80	0.77	0.82	0.87	0.93
SE	0.81	0.88	0.73	0.73	0.33	0.33	0.74	0.81	0.51	0.59	0.60	0.75	0.62	0.64	0.67	0.70	0.42	0.42
SP	0.65	0.71	0.63	0.77	0.81	0.90	0.59	0.65	0.81	0.93	0.56	0.59	0.74	0.86	0.80	0.86	0.89	0.96
MCC	0.47	0.61	0.17	0.26	0.08	0.16	0.22	0.30	0.25	0.49	0.06	0.12	0.34	0.49	0.42	0.52	0.20	0.33
BA	0.73	0.80	0.68	0.75	0.57	0.61	0.67	0.73	0.66	0.76	0.58	0.67	0.68	0.75	0.73	0.78	0.65	0.69
AUC	0.80	0.85	0.72	0.79	0.62	0.68	0.70	0.73	0.75	0.81	0.63	0.67	0.73	0.79	0.82	0.86	0.71	0.71
%	1.00	0.62	1.00	0.45	1.00	0.63	1.00	0.73	1.00	0.44	1.00	0.46	1.00	0.48	1.00	0.77	1.00	0.58

6.10. Robustness - Statistics obtained by Y-scrambling:

	PXR_up	PXR_dn	LXR_up	LXR_dn	AHR_up	AHR_dn	NRF2_up	PPARg_up	PPARa_up
MCC	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

6.11. Robustness - Statistics obtained by bootstrap:

6.12. Robustness - Statistics obtained by other methods:

7. External validation - OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Available information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

No

7.4. Data for the dependent variable for the external validation set:

All

7.5. Other information about the external validation set:

See 6.5

7.6. Experimental design of test set:

See 6.5

7.7. Predictivity - Statistics obtained by external validation:

	PXR_up		PXR_dn		LXR_up		LXR_dn		AHR_up		AHR_dn		NRF2_up		PPARg_up		PPARa_up	
	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD
#	235	144	273	120	275	156	227	184	265	118	272	125	216	112	229	171	266	142
P	128	84	17	6	14	8	29	26	33	14	10	3	70	33	54	36	14	8
N	107	60	256	114	261	148	198	158	232	104	262	122	146	79	175	135	252	134
ACC	0.77	0.85	0.71	0.83	0.77	0.88	0.61	0.66	0.80	0.90	0.52	0.59	0.68	0.75	0.75	0.80	0.89	0.92
SE	0.84	0.94	0.76	0.83	0.50	0.38	0.66	0.65	0.52	0.43	0.40	0.67	0.61	0.58	0.72	0.81	0.57	0.50
SP	0.68	0.73	0.70	0.82	0.78	0.91	0.60	0.66	0.84	0.96	0.53	0.59	0.71	0.82	0.75	0.79	0.91	0.95
MCC	0.54	0.70	0.24	0.35	0.15	0.20	0.17	0.23	0.30	0.45	-0.03	0.08	0.31	0.40	0.42	0.52	0.34	0.39
BA	0.76	0.84	0.73	0.83	0.64	0.64	0.63	0.66	0.68	0.70	0.46	0.63	0.66	0.70	0.74	0.80	0.74	0.72
AUC	0.83	0.85	0.79	0.83	0.62	0.55	0.66	0.67	0.70	0.76	0.54	0.82	0.72	0.76	0.81	0.85	0.71	0.71
%	1.00	0.61	1.00	0.44	1.00	0.57	1.00	0.81	1.00	0.45	1.00	0.46	1.00	0.52	1.00	0.75	1.00	0.53

7.8. Predictivity - Assessment of the external validation set:

MCC values were lower in external validation with respect of internal validation for endpoints with highly unbalanced datasets, as in the case of LXR_up and AhR_dn (MCC < 0.30) for chemicals in the AD. A possible explanation for this poor performance can be found in the extreme degree of imbalance of some VS (i.e. less than 10% of active chemicals) that seriously undermines the reliability of statistical indicators.

Statistical analyses were done to identify critical MCC thresholds for reliably evaluating the performance of models on binary datasets with different degree of imbalance. These thresholds correspond to a reasonable minimum predictivity and were defined for each model by imposing a minimum percentage of correctly predicted positive and negative chemicals of 75% (i.e. SE = SP= 75%). Results demonstrated that, given the same percentage of correctly predicted active and inactive compounds, very unbalanced datasets are linked to lower MCC values. Table below shows which models overcome predictivity thresholds with respect of the degree of unbalance of datasets.

	PXR_up	PXR_dn	LXR_up	LXR_dn	AHR_up	AHR_dn	NRF2_up	PPARg_up	PPARa_up
#	581	482	683	657	463	491	411	699	608
P	322	26	36	70	51	16	110	151	26
N	259	456	647	587	412	475	301	548	582
MCC	0.61	0.26	0.16	0.30	0.49	0.12	0.49	0.52	0.33
MCC75	0.50	0.26	0.25	0.34	0.34	0.20	0.46	0.43	0.24
Valid?	Y	Y	N	Y/N	Y	N	Y	Y	Y
#	144	120	156	184	118	125	112	171	142
P	84	6	8	26	14	3	33	36	8
N	60	114	148	158	104	122	79	135	134
MCC	0.70	0.35	0.20	0.23	0.45	0.08	0.40	0.52	0.39
MCC75	0.49	0.29	0.25	0.39	0.37	0.15	0.47	0.42	0.26
Valid?	Y	Y	N	N	Y	N	N	Y	Y

7.9. Comments on the external validation of the model:

8. Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model:

Not provided

8.2. A priori or a posteriori mechanistic interpretation:

8.3. Other information about the mechanistic interpretation:

9. Miscellaneous information

9.1. Comments:

9.2. Bibliography:

[1] Gadaleta, D., Manganelli, S., Roncaglioni, A., Toma, C., Benfenati, E., Mombelli, E. (2018). QSAR modelling of ToxCast assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. Journal of Chemical Information and Modelling, submitted manuscript.

[2] Judson, R.; Houck, K.; Martin, M.; Richard, A. M.; Knudsen, T. B.; Shah, I.; Little, S.; Wambaugh, J.; Woodrow Setzer, R.; Kothya, P.; Phuong, J.; Filer, D.; Smith, D.; Reif, D.; Rotroff, D.; Kleinstreuer, N.; Sipes, N.; Xia, M.; Huang, R.; Crofton, K.; Thomas, R. S., Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity on *In vitro* Assay Activity Across a Diverse Chemical and Assay Space.Toxicol. Sci. 2016, 152, 323-339.

[3] Romanov, S., Medvedev, A., Gambarian, M., Poltoratskaya, N., Moeser, M., Medvedeva, L., & Makarov, S. (2008). Homogeneous reporter system enables quantitative functional assessment of multiple transcription factors. Nature Methods, 5(3), 253.

[4] Martin, M. T., Dix, D. J., Judson, R. S., Kavlock, R. J., Reif, D. M., Richard, A. M., & Gambarian, M. (2010). Impact of environmental chemicals on key transcription regulators and correlation to toxicity endpoints within EPA's ToxCast program. Chemical research in toxicology, 23(3), 578-590.

[5] Liaw, A.; Wiener, M., Classification and Regression by RandomForest. R News 2002, 2, 18-22.

9.3. Supporting information:

Training set(s)Test set(s)Supporting information

4) Downsampling-based Random Forest with VSURF-based feature selection

QMRF identifier (JRC Inventory):To be entered by JRC

QMRF Title: QSARs for predicting up and down regulation of transcription factors as MIEs of hepatic steatosis (4)

Printing Date:26-giu-2018

1. QSAR identifier

1.1. QSAR identifier (title):

[1] BRF for predicting up regulation of pregnane X receptor (PXR)

[2] BRF RF for predicting down regulation of pregnane X receptor (PXR)

[3] BRF for predicting up regulation of liver X receptor (LXR)

[4] BRF for predicting down regulation of liver X receptor (LXR)

[5] BRF for predicting up regulation of Aryl hydrocarbon receptor (AhR)

[6] BRF for predicting down regulation of Aryl hydrocarbon receptor (AhR)

[7] BRF for predicting up regulation of Nuclear factor (erythroid-derived 2)-like 2 (Nrf2)

[8] BRF for predicting down regulation of Peroxisome proliferator-activated receptors alpha (PPARα)

[9] BRF for predicting down regulation of Peroxisome proliferator-activated receptors gamma (PPARγ)

1.2. Other related models:

1.3. Software coding the model:

KNIME (v3.4).

2. General information

2.1. Date of QMRF:

26 June 2018

2.2. QMRF author(s) and contact details:

Domenico Gadaleta; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; domenico.gadaleta@marionegri.it

2.3. Date of QMRF update(s):

2.4. QMRF update(s):

2.5. Model developer(s) and contact details:

[1] Domenico Gadaleta; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; domenico.gadaleta@marionegri.it

[2] Serena Manganelli; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; serena.manganelli@marionegri.it

[3] Cosimo Toma; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; cosimo.toma@marionegri.it

[4] Alessandra Roncaglioni; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; alessandra.roncaglioni@marionegri.it

[5] Emilio Benfenati; IRCCS - Istituto di Ricerche Farmacologiche Mario Negri; emilio.benfenati@marionegri.it

[6] Enrico Mombelli; Institut National de l'Environnement Industriel et des Risques (INERIS); enrico.mombelli@ineris.fr

2.6. Date of model development and/or publication:

2018

2.7. Reference(s) to main scientific papers and/or software package:

[1] Gadaleta, D., Manganelli, S., Roncaglioni, A., Toma, C., Benfenati, E., Mombelli, E. (2018). QSAR modelling of ToxCast assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. Journal of Chemical Information and Modelling, submitted manuscript.

[2] Judson, R.; Houck, K.; Martin, M.; Richard, A. M.; Knudsen, T. B.; Shah, I.; Little, S.; Wambaugh, J.; Woodrow Setzer, R.; Kothya, P.; Phuong, J.; Filer, D.; Smith, D.; Reif, D.; Rotroff, D.; Kleinstreuer, N.; Sipes, N.; Xia, M.; Huang, R.; Crofton, K.; Thomas, R. S., Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity on *In vitro* Assay Activity Across a Diverse Chemical and Assay Space. Toxicol. Sci. 2016, 152, 323-339.

[3] Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kötter, T.; Meinl, T.; Ohl, P.; Sieb, C.; Thiel, K.; Wiswedel, B., KNIME: The Konstanz Information Miner. In Data Analysis, Machine Learning and Applications, Preisach, C.; Burkhardt, H.; Schmidt-Thieme, L.; Decker, R., Eds.; Springer Berlin Heidelberg: Berlin, Heidelberg, 2008; pp 319-326.

[4] Genuer, R.; Poggi, J. M.; Tuleau-Malot, C., VSURF: An R Package for Variable Selection Using Random Forests. The R Journal 2015, 7, 19-33.

2.8. Availability of information about the model:

All the information about the model are reported in the reference publications (see 2.7).

2.9. Availability of another QMRF for exactly the same model:

3. Defining the endpoint - OECD Principle 1

3.1. Species: Human

3.2. Endpoint:

QMRF 6. Other QMRF 6. 6. Other

3.3. Comment on endpoint:

Up and/or down regulation of activity the following transcription factors:

[1] Pregnane X receptor (PXR) [GeneSymbol:NR1I2 | GeneID:8856 | Uniprot_SwissProt_Accession: O75469], up and down regulation

[2] Liver X recetor (LXR) [GeneSymbol:NR1H2 & NR1H3 | GeneID:7276 & 10062 |

Uniprot_SwissProt_Accession: P55055 & Q13133], up and down regulation [3] Aryl hydrocarbon receptor (AhR) [GeneSymbol: AHR | GeneID: 196 | Uniprot_SwissProt_Accession: P35869], up and down regulation

[4] Nuclear factor (erythroid-derived 2)-like 2 (Nrf2) [GeneSymbol: NFE2L2 | GeneID: 4780 | Uniprot_SwissProt_Accession: Q16236], up regulation

[5] Peroxisome proliferator-activated receptors alpha (PPARα) [GeneSymbol: PPARA | GeneID: 5465 | Uniprot_SwissProt_Accession: Q07869], up regulation

[6] Peroxisome proliferator-activated receptors gamma (PPARγ) [GeneSymbol: PPARA | GeneID: 5468 | Uniprot_SwissProt_Accession: P37231], up regulation

3.4. Endpoint units:

Adimensional

3.5. Dependent variable:

Nine endoints. Categorical (1 for positive, 0 for negative). pAC50 values greater than zero are actives (1), pAC50 values equal to zero are inactives (0). If at least one of the assays considered for each endpoint were active, the sample was flagged as active.

3.6. Experimental protocol:

Assays from ToxCast considered for each endpoint are the following:

[1] PXR up regulation: ATG_PXR_TRANS_up (AEID: 135), and ATG_PXRE_CIS_up (AEID: 103)

[2] PXR down regulation: ATG_PXR_TRANS_dn (AEID: 1475), and ATG_PXRE_CIS_dn (AEID: 1474)

[3] LXR up regulation: ATG_LXRa_TRANS_up (AEID: 125), ATG_LXRb_TRANS_up (AEID: 126), ATG_DR4_LXR_CIS_up (AEID: 70)

[4] LXR down regulation: ATG_LXRa_TRANS_dn (AEID: 1443), ATG_LXRb_TRANS_dn (AEID: 1444), ATG_DR4_LXR_CIS_dn (AEID: 1412)

[5] Ahr up regulation: ATG_Ahr_CIS_up (AEID: 63)

[6] Ahr down regulation: ATG_Ahr_CIS_dn (AEID: 1400)

[7] NrF2 up regulation: ATG_NRF2_ARE_CIS_up (AEID: 97)

[8] PPARα up regulation: ATG_PPARa_TRANS_up (AEID: 132)

[9] PPARγ up regulation: ATG_PPARg_TRANS_up (AEID: 134)

Attagene (ATG) assays are cell-based, multiplexed-redout assays that uses HepG2, a human liver cell line, with measurements taken at 24 hour after chemical dosing in 24-well plate.

These assays are designed to make measurements of mRNA induction, a form of inducible reporter, as detected with fluorescence intensity signals by Reverse transcription polymerase chain reaction (RT-PCR) and Capillary electrophoresis technology.

Changes to fluorescence intensity signals are indicative of inducible changes in transcription factor activity. This is quantified by the level of mRNA reporter sequence unique to:

The cis-acting reporter gene response element which are responsive of an endogenous human receptor subfamily (CIS assays);

The transfected trans-acting reporter gene and exogenous transcription factor GAL4, which are responsive of a given human receptor isoform (TRANS assays).

Further info on the assays: http://www.attagene.com/technology.php

3.7. Endpoint data quality and variability:

Experimental data used in this work were isolated from a collection of 24 *in vitro* HTS assays from the ToxCast program, executed by Attagene Inc. (RTP, NC), under contract to the U.S. EPA (Contract Number EP-W-07-049). During this program, several experiments evaluated the impact of more than 8,000 chemicals on the previously described TFs involved in the MIE of steatosis AOP.

For approximately half the chemicals tested during the ToxCast project, cytotoxicity was observed in the range of concentrations tested. Thus, a significant proportion of measured activities may represent a false positive response caused by assay interference process linked to a cytotoxicity-related 'burst' of activities (Judson *et al.*, 2006, Toxicol. Sci. 2016, 152, 323-339).

The presence of possible false negatives was also reported. The volatility of particular chemical categories (e.g., solvent chemicals) included in ToxCast or the low solubility may explain their general lack of significant effect.

4. Defining the algorithm - OECD Principle 2

4.1. Type of model:

Random Forest (RF)

4.2. Explicit algorithm:

Random Forest was derived based on undersampling of the training set, i.e. random deletion of the most represented class (i.e. negative chemicals) until both classes were equal in number. This approach generated a dataset more suitable for treatment with classical machine learning methods. RF implemented in KNIME was used to derive the undersampling based model. The mtry value was the one provided by default in KNIME. The number of trees was selected according to indications given by VSURF. Indeed, the package enables the identification of best descriptor subsets as a function of the number of trees. Table indicates the number of trees for each model:

PXR		LXR		AhR		NrF2	PPARγ	PPARα
up	dn	up	dn	up	dn	up	up	up
101	51	51	101	101	51	51	51	101

4.3. Descriptors in the model:

[1] PXR_up: ChiA_Dz(Z), Eta_alpha_A, GGI3, MAXDN, MLOGP2, P_VSA_ppp_D, SdssC, P_VSA_ppp_L, SpMax_B(s), SpMax4_Bh(m), SpMax_B(m), SpMin2_Bh(m), SpPosA_B(v), Wap, X5v

[2] PXR_dn: GATS1i, P_VSA_ppp_L, SpPosA_B(p), SpMax1_Bh(p), O%, CATS2D_06_LL, ATSC2e, CATS2D_07_LL

[3] LXR_up: GATS1e, DLS_02, CIC2, TPSA(Tot), SM1_Dz(p), D/Dtr05

[4] LXR_dn: MLOGP2, F01[C-C], CATS2D_04_LL, SIC0, SaasC, P_VSA_e_3

[5] AhR_up: ARR, H-046, GATS1p, MATS1i, F03[N-O], Eta_F_A, MATS2s, P_VSA_m_2

[6] AhR_dn: GATS2e, P_VSA_ppp_L, PHI, B09[C-N], SpMin1_Bh(s), ATSC1s

[7] NrF2_up: P_VSA_ppp_L, SpMax1_Bh(v), MLOGP, MATS1p, ATSC2s, nR=Cp, GATS6e, Eig02_AEA(bo), F10[C-N]

[8] PPARg_up: P_VSA_ppp_L, MLOGP, MLOGP2, PCR, TI2_L, P_VSA_v_3, GGI6, Mi, SpMin2_Bh(e), SM08_AEA(ed), GATS1p, SpMin2_Bh(s), P_VSA_i_4, SpMax5_Bh(m), H-047, nHDon

[9] PPARa_up: GATS8i, rGes, Chi_Dz(Z), GATS5e, LOC, CATS2D_06_DL, ATSC4m, SM04_EA(ed)

4.4. Descriptor selection:

Automated. Use of VSURF (https://cran.r-project.org/web/packages/VSURF/VSURF.pdf) R package.

4.5. Algorithm and descriptor generation:

Descriptors were pruned by constant and semi-constant vales (i.e. standard deviation < 0.01), then if a couple of descriptors was characterised by an absolute pair correlation greater than 90%, the descriptor with the highest pair correlation with all the other descriptors was removed.

Optimal subsets of descriptors for modelling were obtained with the R package VSURF. The algorithm consists in a three step variable selection based on the logic underpinning the random forest (RF) algorithm (i.e. permutation importance and out-of-bag error). The first step eliminates irrelevant descriptors according to the permutation-based RF score of importance and a user-defined threshold. The second step finds important descriptors closely related to the response variable (interpretation step) and the third step (prediction step) identifies a sufficient parsimonious set of important descriptors leading to a good prediction of the response variables. The VSURF selection procedure was carried out as a function of a number of trees ranging from 25 to 251, then the pool of descriptors returning the lowest internal error was retained.

4.6. Software name and version for descriptor generation:

Dragon v7.0.8

Calculation of several sets of molecular descriptors from molecular geometries (topological, geometrical, WHIM, 3D-MoRSE, molecular profiles, etc.)

Kode srl. Via Nino Pisano, 14 56122 Pisa (PI) - Italy, info@kode-solutions.net www.kodesolutions.net

https://chm.kode-solutions.net/products_dragon.php

4.7. Chemicals/Descriptors ratio:

Not relevant for Random Forests

5. Defining the applicability domain - OECD Principle 3

5.1. Description of the applicability domain of the model:

[1] Threshold on confidence of RF

[2] Threshold on similarity of nearest neighbors of the training set

5.2. Method used to assess the applicability domain:

ADs were optimised for each model by fine-tuning the two metrics with respect to predictivity:

[1] For the first AD criterion we estimated the percentage of trees within the RF yielding the same prediction (i.e. confidence). A confidence threshold (Tc) was implemented in KNIME and gradually incremented by 0.05, from 0.55 to 0.75. Chemicals with confidence lower than this threshold were considered as outside the model AD.

[2] The second AD criterion took account of the structural domain of the model. This was achieved by evaluating the degree of structural similarity of a given compound to those included within the TS. A distance matrix containing Euclidean distances for each pair of chemicals in the TS was calculated, then for each TS compound the mean distance from its first k neighbors was calculated. TS chemicals were then sorted on the basis of these distances and the value corresponding to a given percentile of the distribution of distances was used as a threshold (TD) beyond which chemicals were excluded from the AD. For the external validations, the same procedure was repeated calculating the prediction confidence and the distances of each VS chemical from their neighbors within the TS, then TD was used to identify chemicals outside of AD. For the present work, we adopted the Euclidean distance calculated on the scaled and centered descriptors used by the models as a similarity criterion; values assigned to k were 1 and 5; values assigned to TD were those corresponding to the 100th, the 97.5th, the 95th and the 90th percentiles of the TS distance distributions.

The best values for parameters were selected based on performance and coverage (i.e., percentage of predictions in the AD) in 10-fold-cross-validation. The final values are:

	PXR		LXR		AhR		NrF2	PPARγ	PPARα
	up	dn	up	dn	up	dn	up	up	up
T_D	0.60	0.75	0.70	0.75	0.70	0.65	0.70	0.65	0.60
T_C	100th	95th	90th	95th	97.5th	97.5th	100th	100th	90th
MN	1	1	5	5	5	1	1	1	1

5.3. Software name and version for applicability domain assessment:

KNIME (Data Analysis, Machine Learning and Applications. Studies in Classification, Data Analysis, and Knowledge Organization); Prof. Dr. Michael Berthold, Michael.Berthold@uni-konstanz.de

https://www.knime.com/

5.4. Limits of applicability:

The model is not applicable on inorganic chemicals and those including unusual elements (i.e., different from C, O, N, S, Cl, Br, F, I). Salts can be predicted only if stripped of the counterion and converted to the neutralised form.

6. Internal validation - OECD Principle 4

6.1. Availability of the training set:

Yes

6.2. Available information for the training set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

6.3. Data for each descriptor variable for the training set:

Not available

6.4. Data for the dependent variable for the training set:

All

6.5. Other information about the training set:

Datasets for each endpoint were randomly divided into a training set (TS, 80% of the original dataset) and a validation set (VS, 20% of the original dataset) comprising the same proportion of active and inactive chemicals as the original dataset. The table below reports the number of chemicals in TS and VS for each of the modelled endpoint.

TF	Inactives (full database)	Actives (full database)	% Actives	TS	TS_{us}	VS
PXR_up	529	640	55	934	934	235
PXR_dn	1269	83	6	1079	132	273
LXR_up	1296	68	5	1089	108	275
LXR_dn	990	141	12	904	224	227
AhR_up	1148	162	12	1045	258	265
AhR_dn	1301	50	4	1079	80	272
NRF2_up	723	346	32	853	552	216
PPARγ_up	871	266	23	908	424	229
PPARα_u	1259	64	5	1057	100	266

6.6. Pre-processing of data before modelling:

Data were retrieved from the oldstyle_neg_log_ac50_Matrix_151020.csv file, downloaded from ftp://newftp.epa.gov/comptox/High_Throughput_Screening_Data/Summary_Files.

For classification purposes, for each assay chemicals with zero values for a given assay were considered as inactive, while chemicals with a continuous pAC50 value were considered active.

Results from TRANS-assays were considered if specific isoforms of a TF listed among the AOPs for steatosis (e.g., PPARα, PPARγ), while CIS-assays or a combination of CIS- and TRANS-assays were used if TF isoforms were not specified. In the latter case, CIS- and TRANS- outputs were combined. A chemical was labelled as active if it was active in at least one assay and inactive if it was inactive in both types of assay.

Only chemicals exceeding 90% purity were retained, while chemicals associated to lower purity, other anomalies (e.g. withdrawn chemicals) or not yet analysed were not included.

The structures were checked by removing inorganic chemicals and mixtures, correcting inaccurate SMILES codes with the help of chemical databases, i.e. ChemSpider and ChemIDplus and neutralizing salts.

An in-house software was used to identify and remove duplicates. For a given set of duplicated structures, if their experimental activities were identical, then only one compound was kept. If their experimental properties were different, both the chemicals were removed.

A Python script executing the MolVS standardiser (based on RDkit libraries) was written to obtain canonical tautomers. Canonical SMILES were coded using the istMolBase software based on CDK libraries.

A z-score (Eq. 1) that was assigned to each chemical-assay combination (Judson *et al.*, 2006, Toxicol. Sci. 2016, 152, 323-339):

$$Z(chemical\,,assay) = \frac{-\log AC_{50}(chemical\,,assay) - median[-\log AC_{50}(\text{chemical}\,,\text{cytotoxity})]}{global\ cytotoxicity\ MAD} \quad (1)$$

Conversely, chemicals associated with low z-scores are more likely to be false positives confounded by cytotoxicity. A z-score threshold of 3 was considered to select only chemicals that can be considered as specifically active.

6.7. Statistics for goodness-of-fit:

6.8. Robustness - Statistics obtained by leave-one-out cross-validation:

6.9. Robustness - Statistics obtained by leave-many-out cross-validation:

Method: 10-fold cross-validation. Results are reported above:

	PXR_up		PXR_dn		LXR_up		LXR_dn		AHR_up		AHR_dn		NRF2_up		PPARg_up		PPARa_up	
	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD
#	934	747	132	76	108	61	224	97	258	123	80	53	552	255	424	301	100	72
P	512	407	66	37	54	35	112	45	129	62	40	24	276	121	212	159	50	38
N	422	340	66	39	54	26	112	52	129	61	40	29	276	134	212	142	50	34
ACC	0.74	0.78	0.83	0.89	0.72	0.82	0.69	0.78	0.73	0.83	0.75	0.85	0.69	0.76	0.76	0.81	0.78	0.82
SE	0.79	0.85	0.83	0.92	0.72	0.83	0.69	0.73	0.77	0.84	0.75	0.88	0.67	0.79	0.75	0.82	0.76	0.82
SP	0.68	0.69	0.82	0.87	0.72	0.81	0.70	0.83	0.70	0.82	0.75	0.83	0.71	0.75	0.77	0.80	0.80	0.82
MCC	0.47	0.55	0.65	0.79	0.44	0.63	0.38	0.56	0.47	0.66	0.50	0.70	0.38	0.53	0.52	0.62	0.56	0.64
BA	0.73	0.77	0.83	0.90	0.72	0.82	0.69	0.78	0.73	0.83	0.75	0.85	0.69	0.77	0.76	0.81	0.78	0.82
AUC	0.80	0.82	0.86	0.87	0.77	0.83	0.73	0.81	0.80	0.87	0.80	0.82	0.75	0.80	0.82	0.86	0.79	0.83
%	1.00	0.80	1.00	0.58	1.00	0.56	1.00	0.43	1.00	0.48	1.00	0.66	1.00	0.46	1.00	0.71	1.00	0.72

6.10. Robustness - Statistics obtained by Y-scrambling:

	PXR_up	PXR_dn	LXR_up	LXR_dn	AHR_up	AHR_dn	NRF2_up	PPARg_up	PPARa_up
MCC	0.00	-0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00

6.11. Robustness - Statistics obtained by bootstrap:

6.12. Robustness - Statistics obtained by other methods:

7. External validation - OECD Principle 4

7.1. Availability of the external validation set:

Yes

7.2. Available information for the external validation set:

CAS RN: Yes

Chemical Name: Yes

Smiles: Yes

Formula: No

INChI: No

MOL file: No

7.3. Data for each descriptor variable for the external validation set:

No

7.4. Data for the dependent variable for the external validation set:

All

7.5. Other information about the external validation set:

See 6.5

7.6. Experimental design of test set:

See 6.5

7.7. Predictivity - Statistics obtained by external validation:

	PXR_up		PXR_dn		LXR_up		LXR_dn		AHR_up		AHR_dn		NRF2_up		PPARg_up		PPARa_up	
	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD	all	AD
#	235	177	273	132	275	115	227	102	265	114	272	175	216	101	229	168	266	180
P	128	102	17	7	14	5	29	11	33	20	10	8	70	27	54	41	14	11
N	107	75	256	125	261	110	198	91	232	94	262	167	146	74	175	127	252	169
ACC	0.78	0.84	0.63	0.70	0.56	0.53	0.63	0.76	0.66	0.75	0.53	0.51	0.66	0.75	0.73	0.78	0.70	0.73
SE	0.87	0.92	0.65	1.00	0.64	0.80	0.59	0.73	0.55	0.65	1.00	1.00	0.67	0.78	0.81	0.88	0.64	0.73
SP	0.67	0.73	0.63	0.69	0.56	0.52	0.63	0.77	0.67	0.78	0.52	0.49	0.65	0.74	0.70	0.75	0.71	0.73
MCC	0.55	0.68	0.14	0.32	0.09	0.13	0.15	0.34	0.15	0.35	0.19	0.20	0.30	0.47	0.45	0.55	0.17	0.24
BA	0.77	0.83	0.64	0.84	0.60	0.66	0.61	0.75	0.61	0.71	0.76	0.74	0.66	0.76	0.76	0.81	0.67	0.73
AUC	0.84	0.86	0.71	0.84	0.61	0.62	0.65	0.75	0.65	0.74	0.79	0.74	0.71	0.80	0.82	0.84	0.69	0.69
%	1.00	0.75	1.00	0.48	1.00	0.42	1.00	0.45	1.00	0.43	1.00	0.64	1.00	0.47	1.00	0.73	1.00	0.68

7.8. Predictivity - Assessment of the external validation set:

MCC values were lower in external validation with respect of internal validation for endpoints with highly unbalanced datasets, as in the case of LXR_up and AhR_dn (MCC < 0.30) for chemicals in the AD. A possible explanation for this poor performance can be found in the extreme degree of imbalance of some VS (i.e. less than 10% of active chemicals) that seriously undermines the reliability of statistical indicators.

Statistical analyses were done to identify critical MCC thresholds for reliably evaluating the performance of models on binary datasets with different degree of imbalance. These thresholds correspond to a reasonable minimum predictivity and were defined for each model by imposing a minimum percentage of correctly predicted positive and negative chemicals of 75% (i.e. SE = SP=75%). Results demonstrated that, given the same percentage of correctly predicted active and inactive compounds, very unbalanced datasets are linked to lower MCC values. Table below shows which models overcome predictivity thresholds with respect of the degree of unbalance of datasets.

	PXR_up	PXR_dn	LXR_up	LXR_dn	AHR_up	AHR_dn	NRF2_up	PPARg_up	PPARa_up
#	747	76	61	97	123	53	255	301	72
P	407	37	35	45	62	24	121	159	38
N	340	39	26	52	61	29	134	142	34
MCC	0.55	0.79	0.63	0.56	0.66	0.7	0.53	0.62	0.64
MCC75	0.50	0.50	0.51	0.50	0.51	0.51	0.51	0.50	0.53
Valid?	Y	Y	Y	Y	Y	Y	Y	Y	Y
#	177	132	115	102	114	175	101	168	180
P	102	7	5	11	20	8	27	41	11
N	75	125	110	91	94	167	74	127	169
MCC	0.68	0.32	0.13	0.34	0.35	0.2	0.47	0.55	0.24
MCC75	0.50	0.23	0.25	0.32	0.41	0.23	0.45	0.45	0.26
Valid?	Y	Y	N	Y	N	Y/N	Y	Y	Y/N

7.9. Comments on the external validation of the model:

8. Providing a mechanistic interpretation - OECD Principle 5

8.1. Mechanistic basis of the model:

Not provided

8.2. A priori or a posteriori mechanistic interpretation:

8.3. Other information about the mechanistic interpretation:

9. Miscellaneous information

9.1. Comments:

9.2. Bibliography:

[1] Gadaleta, D., Manganelli, S., Roncaglioni, A., Toma, C., Benfenati, E., Mombelli, E. (2018). QSAR modelling of ToxCast assays Relevant to the Molecular Initiating Events of AOPs Leading to Hepatic Steatosis. Journal of Chemical Information and Modelling, submitted manuscript.

[2] Judson, R.; Houck, K.; Martin, M.; Richard, A. M.; Knudsen, T. B.; Shah, I.; Little, S.; Wambaugh, J.; Woodrow Setzer, R.; Kothya, P.; Phuong, J.; Filer, D.; Smith, D.; Reif, D.; Rotroff, D.; Kleinstreuer, N.; Sipes, N.; Xia, M.; Huang, R.; Crofton, K.; Thomas, R. S., Editor's Highlight: Analysis of the Effects of Cell Stress and Cytotoxicity on *In vitro* Assay Activity Across a Diverse Chemical and Assay Space. Toxicol. Sci. 2016, 152, 323-339.

[3] Romanov, S., Medvedev, A., Gambarian, M., Poltoratskaya, N., Moeser, M., Medvedeva, L., & Makarov, S. (2008). Homogeneous reporter system enables quantitative functional assessment of multiple transcription factors. Nature Methods, 5(3), 253.

[4] Martin, M. T., Dix, D. J., Judson, R. S., Kavlock, R. J., Reif, D. M., Richard, A. M., & Gambarian, M. (2010). Impact of environmental chemicals on key transcription regulators and correlation to toxicity end points within EPA's ToxCast program. Chemical research in toxicology, 23(3), 578-590.

9.3. Supporting information:

Training set(s)Test set(s)Supporting information

8. Dempster-Shafer theory for combining evidence and estimating uncertainty

Introduction

In vitro and *silico* methods play an important part in assessing the safety of chemical. In order to make unbiased decisions on incomplete data from a number of different sources, e g. assay outcomes or predictions from *in silico* models, the reliability and uncertainty of the various sources need to be taken into account.

By far the most common method used for combining evidence from multiple source is the consensus approach; e.g., if two sources predict outcome A and one source predict outcome B, then outcome A is reported as the consensus prediction. The consensus method is easy to apply but, naively, assumes that all models are equally reliable. Also, the uncertainty of each source is not taken into account. In order to account for these limitations Dempster-Shafer theory (DST) [1-2], is used for providing an unbiased decision based on the quality and reliability of the sources.

Method

DST is an extension of generalised Bayesian statistical inference in which evidence can be associated with multiple events.

The following example is used to illustrate the general theory and algorithms of DST:

Binary case (Y) with an "event" being Toxic(T) or non-Toxic (N) is investigated:

In a traditional probability approach - the probabilities are additive: p(Y) ≡ p(T or N) = p(T) + p(N) DST provides a mechanism to take uncertainty due into account and replaces traditional probability functions with *belief* and *plausibility* functions.

Thus the set of outcomes of Y is P(Y) = {{y1},{y2},{y1,y2},∅}

where ∅ denotes the empty set.

In this way *belief* and *plausibility* can be viewed as lower and upper bounds, respectively, of the probability for the outcome.

DST uses *probability mass functions* (m):

$0 \leqslant m(f_i) \leqslant 1$

$m(\emptyset) = 0$

$\Sigma m(f_i) = 1$

Having a binary variable Y={y1,y2}, (T or N) the probability masses can be assigned to:

{y1} and {y2} and to {y1,y2}:

$\Sigma m(f_i) = 1 = m(\{y1\}) + m(\{y2\}) + m(\{y1,y2\})$

The probability masses m({y1}) and m({y2}) is the proportion of the overall belief with respect to outcomes y_1 and y_2, respectively, and m({y1,y2}) is the proportion of the overall belief not committed to y_1 but also not ascribed to y_2 as well as not committed to y_2 but also

not ascribed to y_I. Thus m({y1,y2}) quantifies the level of uncertainty in the system under investigation.

Belief (Bel) and plausibility (Pls) is defined for the 2 outcomes T and N as:

Bel({T}) = m({T}) Bel({N}) = m({N}

Pls({T}) = m({T}) + m({T,N}) Pls({N}) = m({N}) + m({T,N})

Assay 1 predicts 70% probability for toxicity of the investigated compound.

From experience (validation), the reliability of the assay is estimated to 60%.

This gives the following masses, beliefs and plausibilities:

m({T}) = p(T) * rel(T) = (0.70)*(0.60) = 0.42

40% chance (100 -60) that the test result is unreliable ->

basic probability mass associated with uncertainty: **m({T,N})** = 0.40

Sum of masses must be one: **m({N})** = 1 - 0.42 - 0.40 = 0.18

Beliefs and plausibilities are then calculated to be:

Bel(T) = 0.42 Pls(T) = 0.42 + 0.40 = 0.82

Bel(N) = 0.18 Pls(N) = 0.18 + 0.40 = 0.58

If 2 sources are combined then the probability masses are:

q(fk) = m1(fi)*m2(fj)

q(∅) = m1(∅)*m2(∅)

q(Y) = m1(Y)*m2(Y)

By using Dempster's combination rule and assigning all conflicts (to assays give different results) to the null set, normalizing by dividing by 1-q(∅) to insure mD(∅)=0 as required, to joint basic probability mass m_D is mD(Y)=q(Y)/(1-q(∅)). The same normalisation is performed for q(fk).

For two assays, 1 and 2 the following DST scheme is then constructed:

Assay 1 predicts 70% probability for toxicity
From experience (validation) - estimation of reliability = 60%

Assay 2 predicts 60% probability for toxicity
From experience (validation) - estimation of reliability = 80%

For combining the sources:

Applying the Dempster combination rule

$q(\{T\}) = m1(\{T\})*m2(\{T\}) + m1(\{T\})*m2(\{T,N\}) + m1(\{T,N\})*m2(\{T\})$

$q(\{N\}) = m1(\{N\})*m2(\{N\}) + m1(\{N\})*m2(\{T,N\}) + m1(\{T,N\})*m2(\{N\})$

$q(Y) \equiv q(\{T,N\}) = m1(\{T,N\})*m2(\{T,N\})$

$q(\emptyset) = m1(\{T\})*m2(\{N\}) + m1(\{N\})*m2(\{T\})$

This, in turn, gives the following outcome from the DST analysis.

	Assay1 **70 %** probability for toxicity **30 %** probability for non-toxicity Reliability 60 %	Assay2 **60 %** probability for toxicity **40 %** probability for non-toxicity Reliability 80 %
	m1({T}) = 42 %, m1({N}) = 18 % , m1({T,N}) = 40 %	m2({T}) = 48 %, m1({N}) = 32 % , m1({T,N}) = 20 %
Bel ({T}) = Pls ({T}) =	42 % 40 + 42 = 82 %	48 % 48 + 20 = 62 %
Bel ({N}) = Pls ({N}) =	18 % 18 + 40 = 58 %	32 % 32 + 20 = 52 %

$q(\{T\}) = m1(\{T\})*m2(\{T\}) + m1(\{T\})*m2(\{T,N\}) + m1(\{T,N\})*m2(\{T\})$ (= 0.48)
$q(\{N\}) = m1(\{N\})*m2(\{N\}) + m1(\{N\})*m2(\{T,N\}) + m1(\{T,N\})*m2(\{N\})$ (= 0.22)
$q(Y) \equiv q(\{T,N\}) = m1(\{T,N\})*m2(\{T,N\})$ (= 0.08)
$q(\emptyset) = m1(\{T\})*m2(\{N\}) + m1(\{N\})*m2(\{T\})$ (= 0.22)

Dempster combination rule [$mD(Y)=q(Y)/(1-q(\emptyset))$]

$q(\{T\}) = 0.48/(1-0.22) = 0.615$
$q(\{N\}) = 0.22/(1-0.22) = 0.282$
$q(\{T,N\}) = 0.08/(1-0.22) = 0.103$

Bel ({T}) = Pls ({T}) =	61.5 % 61.5 + 10.3 = 71.8 %
Bel ({N}) = Pls ({N}) =	28.2 % 28.2 + 10.3 = 38.5 %

The results is that the chemical the belief for being toxic and non-toxic is 61.5 % and 28.2 %, respectively, with an uncertainty of 10.3 %. Since the plausibility for the compound is 38.5,

well below 50 %, as well as the belief and plausibility for being toxic is above 50% the compound would be considered as toxic.

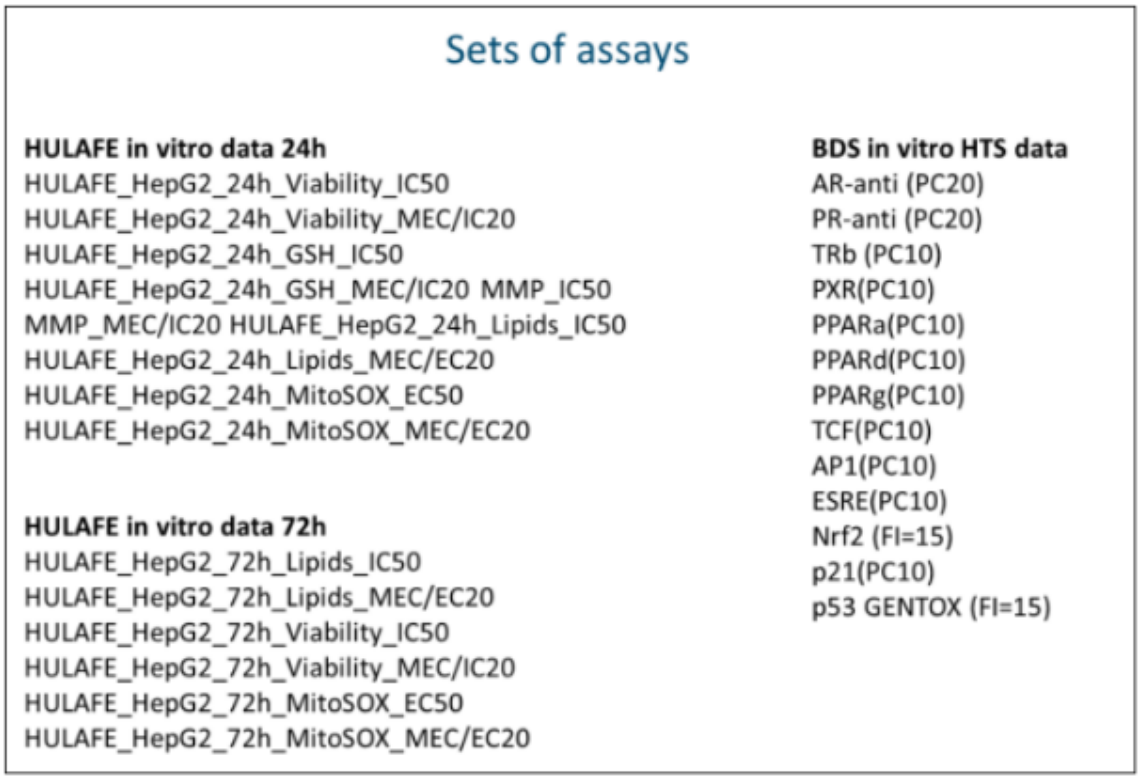
Software

The DST software applied in this study was developed in the EU-ADR project (ICT-215847) and implemented in Java.

To facilitate easier usage of the software several utility programs for data input and analysis were developed using Python.

Read-across using DST

The read-across were performed using 3 sets of assays (BDS HTS assays, HULAFE *in vitro* data 24h and HULAFE *in vitro* data 72h).



Class column (class =1 -> steatotic; class = -1 -> non-steatotic

BDS in vitro HTS data

CAS_2ab	class	AR-ant (PC20)	PR-ant (PC20)	TRb (PC10)	PXR(PC10)	PPARa(PC10)	PPARd(PC10)	PPARg(PC10)	TCF(PC10)	AP1(PC10)	ERE(PC10)	Nrf2 (Fi>15)	p21(PC10)	p53 GENTOX (Fi>15)	set
99661	1	0	1	1	1	1	0	0	1	0	1	0	1	1	source
149575	1	1	1	0	1	1	0	0	1	0	1	1	1	1	source
75989	-1	0	0	0	0	0	0	0	0	0	0	0	0	0	source
88095	-1	0	0	0	0	1	0	0	0	0	0	0	0	0	target

HULAFE in vitro data 24h

CAS	class	HULAFE_Hep G2_24h_Viability_IC50	HULAFE_Hep G2_24h_Viability_MEC/IC20	HULAFE_Hep G2_24h_GSH_IC50	HULAFE_Hep G2_24h_GSH_MEC/IC20	MMP_IC50	MMP_MEC/IC20	HULAFE_Hep G2_24h_Lipids_IC50	HULAFE_Hep G2_24h_Lipids_MEC/EC20	HULAFE_Hep G2_24h_Mito SOX_EC50	HULAFE_Hep G2_24h_Mito SOX_MEC/EC20	set
149575	1	1	1	0	0	0	0	1	1	1	1	source
99661	1	0	0	1	1	0	1	1	1	1	1	source
75989	-1	0	0	0	0	0	0	0	0	0	0	source
88095	-1	0	0	0	0	0	0	0	0	0	0	target

HULAFE in vitro data 72h

HULAFE_Hep G2_72h_CAS	class	HULAFE_Hep G2_72h_Lipids_IC50	HULAFE_Hep G2_72h_Lipids_MEC/EC20	HULAFE_Hep G2_72h_Viability_IC50	HULAFE_Hep G2_72h_Viability_MEC/IC20	HULAFE_Hep G2_72h_Mito SOX_EC50	HULAFE_Hep G2_72h_Mito SOX_MEC/EC20	set		sub-selection
149575	1	0	1	0	0	0	1	source		
99661	1	1	1	1	1	0	1	source		
75989	-1	0	0	0	1	0	1	source		
88095	-1	0	0	0	0	0	0	target		

A leave-out-out cross-validation was performed in order to calculate the positive (PPV) and negative prediction (NPV) value, respectively, for the assays based on the 3 source compounds.

This resulted in the following outcome [posPredAcc (PPV), nedPredAcc (NPV)], balanced accuracy as reliability estimate):

assay_provider	neg class value	posPredAcc	negPredAcc	reliability (Balanced accuracy)	method_name	specificity	sensitivity	selected assays
BDS	-1	1	1	1	ESRE_PC10_	1	1	
BDS	-1	1	1	1	p21_PC10_	1	1	
BDS	-1	1	1	1	p53_GENTOX__FI_15_	1	1	
BDS	-1	1	1	1	PPARa_PC10_	1	1	
BDS	-1	1	1	1	PR-anti__PC20_	1	1	
BDS	-1	1	1	1	PXR_PC10_	1	1	
BDS	-1	1	1	1	TCF_PC10_	1	1	
BDS	-1	1	0.5	0.75	AR-anti__PC20_	1	0.5	
BDS	-1	1	0.5	0.75	Nrf2__FI_15_	1	0.5	
BDS	-1	1	0.5	0.75	TRb__PC10_	1	0.5	
BDS	-1	0	0.333	0.5	AP1_PC10_	1	0	
BDS	-1	0	0.333	0.5	PPARd_PC10_	1	0	
BDS	-1	0	0.333	0.5	PPARg_PC10_	1	0	
HULAFE_24h	-1	1	1	1	HULAFE_HepG2_24h_Lipids_IC50	1	1	
HULAFE_24h	-1	1	1	1	HULAFE_HepG2_24h_Lipids_MEC_EC20	1	1	
HULAFE_24h	-1	1	1	1	HULAFE_HepG2_24h_MitoSOX_EC50	1	1	
HULAFE_24h	-1	1	1	1	HULAFE_HepG2_24h_MitoSOX_MEC_EC20	1	1	
HULAFE_24h	-1	1	0.5	0.75	HULAFE_HepG2_24h_GSH_IC50	1	0.5	
HULAFE_24h	-1	1	0.5	0.75	HULAFE_HepG2_24h_GSH_MEC_IC20	1	0.5	
HULAFE_24h	-1	1	0.5	0.75	HULAFE_HepG2_24h_Viability_IC50	1	0.5	
HULAFE_24h	-1	1	0.5	0.75	HULAFE_HepG2_24h_Viability_MEC_IC20	1	0.5	
HULAFE_72h	-1	1	1	1	HULAFE_HepG2_72h_Lipids_MEC_EC20	1	1	
HULAFE_72h	-1	1	0.5	0.75	HULAFE_HepG2_72h_Lipids_IC50	1	0.5	
HULAFE_72h	-1	1	0.5	0.75	HULAFE_HepG2_72h_Viability_IC50	1	0.5	
HULAFE_72h	-1	0	0.333	0.5	HULAFE_HepG2_72h_MitoSOX_EC50	1	0	sel
HULAFE_72h	-1	0.667	0	0.5	HULAFE_HepG2_72h_MitoSOX_MEC_EC20	0	1	sel
HULAFE_72h	-1	0.5	0	0.25	HULAFE_HepG2_72h_Viability_MEC_IC20	0	0.5	sel

Using the correlations between *in vivo* and *in vitro* outcomes from 3 sets of assays (BDS HTS assays, HULAFE *in vitro* data 24h and HULAFE *in vitro* data 72h), respectively, for the 3 source compounds the *in vivo* predicted outcome for the target compound in all 3 cases (one from each set of assays) is that it is not steatotic.

Results from DST on target compound:

id	belief	plausibility	DST outcome	classification	true_ class	assay set			comment
88095	0	0	low	non-steatotic	non-steatotic	cs1_dst_BDS_selected_assays_binary_excluded_nodata			BDS set
88095	0	0	low	non-steatotic	non-steatotic	cs1_dst_HULAFE_24h_binary_excluded_nodata			HULAFE_24h set
88095	0	0	low	non-steatotic	non-steatotic	cs1_dst_HULAFE_72h_binary_excluded_nodata			HULAFE_72h set
88095	0	0	low	non-steatotic	non-steatotic	HULAFE_HepG2_72h_Lipids_IC50	HULAFE_HepG2_72h_Lipids_MEC/EC20	HULAFE_HepG2_72h_Viability_IC50	HULAFE_72h set, 3 best assays
88095	0	0.833	moderate	?	non-steatotic	HULAFE_HepG2_72h_Viability_MEC/IC20	HULAFE_HepG2_72h_MitoSOX_EC50	HULAFE_HepG2_72h_MitoSOX_MEC/EC20	HULAFE_72h set, 3 worst assays

The DST analysis also shows that using only the 3 best HULAFE assays according to PPV, NPV and reliability provides the same outcome as using all of the assays while deliberately using the 3 worst assays would lead to an inconclusive results due to the fact that the belief is 0 (<0.5) but with a very high uncertainty (plausibility) of 0.833 (> 0.5).

References

1. G. Shafer, A Mathematical Theory Of Evidence, Princeton University Press (1976)
2. A.P. Dempster, Upper and lower probabilities induced by a multivalued mapping, Ann. Math.Stat., 38 (2) (1967), pp. 325-339