Mutation Research

Key Read Across Framework Components and Biology Based Improvements

Nicholas Ball[a], Judith Madden[b], Alicia Paini[f], Miriam Mathea[c],
Andrew David Palmer[C], Saskia Sperber[d], Thomas Hartung[e],
Bennard van Ravenzwaay[d]

a Dow Chemical Company, Horgen, Switzerland

b School of Pharmacy and Bimolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool, UK.

c BASF SE, computational chemistry, Ludwigshafen, Germany

d BASF SE, experimental toxicology and ecology, Ludwigshafen, Germany

e Johns Hopkins University, Center for Alternatives to Animal Testing (CAAT), Baltimore, USA, and University of Konstanz, CAAT-Europe, Germany

f European Commission Joint Research Centre, Ispra, Italy.

Corresponding author: Bennard van Ravenzwaay, BASF SE, Ludwigshafen
bennard.ravenzwaay@basf.com

Abstract

At the 2019 annual meeting of the European Environmental Mutagen and Genomics Society a workshop session related to the use of read across concepts in toxicology was held. The goal of this session was to provide the audience an overview of general read-across concepts. From ECHA's read across assessment framework, the starting point is chemical similarity. There are several approaches and algorithms available for calculating chemical similarity based on molecular descriptors, distance/similarity measures and weighting schemata for specific endpoints. Therefore, algorithms that adapt themselves to the data (endpoint/s) and provide a good ability to distinguish between structural similar and not similar molecules regarding specific endpoints are needed and their use discussed. Toxico-dynamic end points are usually in the focus of read across cases. However, without appropriate attention to kinetics and metabolism such cases are unlikely to be successful. To further enhance the quality of read across cases new approach methods can be very useful. Examples based on a biological approach using plasma metabolomics in rats are given. Finally, with the availability of large data sets of structure activity relationships, in silico tools have been developed which provide hitherto undiscovered information. Automated process is now able to assess the chemical – activity space around the molecule target substance and examples are given demonstrating a high predictivity for certain endpoints of toxicity. Thus, this session provides not only current state of the art criteria for good read across, but also indicates how read-across can be further developed in the near future.

Keywords:

REACH, read-across, in silico, metabolomics, PBPK modelling

Introduction

‘Read-across and grouping’, or ‘read-across’, is one of the most commonly used alternative approaches for data gap filling in registrations submitted under the REACH Regulation. Read-across entails the use of relevant information from analogous substances (the ‘source’ information) to predict properties for the ‘target’ substance(s) under consideration. The conditions under which ‘Read-across and grouping’ can be used to adapt the standard testing regime are listed in Annex XI, 1.5 of the REACH Regulation. It has to be ensured that prediction of a property based on read-across is reliable, can be used for risk assessment and/or classification and labelling, and complies in general with the provisions in REACH for the substance under consideration. Registrants are obligated to consider and, where they can, use appropriate alternative approaches to fulfil applicable REACH information requirements concerning vertebrate animal studies. If read-across which meets the information requirements is applied, unnecessary animal testing may be avoided as there will be no need to carry out one-by-one testing of all their substances to fulfil the information requirements (ECHA RAAF (2015).

Guidance on how to build appropriate read across cases is also given in ECHA’s Read-Across Assessment Framework (RAAF). Using this guidance is theoretically not too difficult, however in practice, experience has shown that only a few solid cases could be built to significantly reduce animal testing for repeated dose studies (systemic toxicity studies with a duration > 4 weeks and reproductive toxicity).

ECHA has attempted to address such issues in a workshop in which new approach methods (encompassing classical in vitro alternative methods, in silico approaches and new technologies) were evaluated for their practical use in regulatory decision making.

ECHA’s Topical Scientific Workshop (19-20 April 2016) addressed the use of data and information from new approach methodologies (NAMs) to support regulatory decisions for the use of chemical substances. An international audience considered three themes representing the use of NAMs for read-across (Theme 1), for screening and prioritisation (Theme 2) and for future prospects (Theme 3). The main deliberations and conclusions of the workshop are summarised in this document.

NAMs were taken in a broad context to include in silico approaches, in chemico and in vitro assays, as well as the inclusion of information from the exposure of chemicals in the context of hazard assessment. They also include a variety of new testing tools, such as “high-throughput screening” and “high-content methods” e.g. genomics, proteomics, metabolomics; as well as some “conventional” methods that aim to improve understanding of toxic effects, either through improving toxicokinetic or toxicodynamic knowledge for substances.

Three read-across case studies were presented, including an evaluation of the read-across, and the contributions of NAMs to reduce uncertainty, using ECHA’s Read-across assessment framework (RAAF).

NAMs were found to support read-across, especially by providing information on toxicodynamics, which increased confidence in mechanistic hypotheses and justification. However, the NAM approaches considered in this workshop were found to be less useful to provide evidence on toxicokinetics to support a read-across argument. NAMs were also shown to be applied in a variety of scenarios for screening and prioritisation with examples from various regions. The future prospects for the use of NAMs were outlined through presentations on current and anticipated practice. The workshop recognised the usefulness of the NAMs for a number of regulatory uses.

Three years later, progress and knowledge about read-across in general and the use of NAMs in this context has advanced, but communication has been rather limited. Therefore, a workshop was organized by ECETOC (the European Centre for Ecotoxicology and Toxicology of Chemicals) – a Brussels based science institute, at the 2019 EEMGS (European Environmental Mutagen and Genomics Society) to promote communication and discussion of read across concepts to reduce animal testing.

The workshop consisted of 5 presentations and a concluding discussion, as shown below.

Session III ECETOC Workshop Read across

Nicholas Ball: Read across – a regulatory perspective based on the RAAF

Judith Madden: Kinetics and PBPK modelling as a component for read across

Miriam Mathea: Structural similarities and molecular design

Saskia Sperber: The use of metabolomics to improve the quality of read across

Thomas Hartung: Big Data In silico tools for read across

The essentials of the individual presentations are summarized in the following.

1. Read across – a regulatory perspective based on the read across assessment framework (RAAF), with particular emphasis on Mutagenicity

Why use read across for genotoxicity / mutagenicity ?

Nicholas Ball presented how the assessment of genotoxicity and the potential for mutagenicity using a tiered approach to testing, beginning with in silico assessment of potential structural alerts for known mechanisms of DNA interaction can be done. It moves through a suite of in vitro assays which typically comprise bacterial mutagenicity, mammalian cell clastogenicity and mutagenicity, and finally, and if warranted, a broader array of in vivo assays to determine the specific genotoxic/mutagenic effects being produced and their potential relevance to humans. Beyond this, it may be also necessary to follow up with more extensive testing to further characterise the outcome associated with mutagenicity, such as assessing carcinogenicity or transgenerational effects on reproduction [1, 2] Consequently, the assessment of genotoxic and mutagenic potential can evolve into a complex and resource intensive (including animal use) exercise. In addition, no assay is perfect and therefore there is also the possibility that a false positive or negative assay triggers an assessment and further testing which is ultimately inappropriate. It is therefore important to utilize as much information as possible as part of a weight of evidence when determining how to characterize this endpoint and focus effort on those substances where it is most appropriate.

One way to achieve this is to group structurally similar substances together and utilize read-across of data from one or more members of the group to the others. This provides an opportunity to make the most out of the data you have access to and potentially limits the need for testing to only a subset of group members. It also offers a way to interrogate positive and negative results in the context of the broader group of substances to assess whether the results of an assay are reliable or perform a more in-depth mechanism of action assessment without generating data on every substance. As such the use of read-across can play an important role in the assessment of a substance’s genotoxicity and mutagenicity.

One of the key criteria when forming a category of substances is the members should be structurally similar. As such, the use of read-across within categories of substances is well suited to assess mutagenicity due to the strong link between chemical structure and ability to interact with DNA. The close association between chemical structure (presence or absence of ‘active functional groups) and the ability to interact with DNA is also why the prediction of genotoxicity using QSAR tools is so well developed versus other endpoints with a more complex mechanism of action where chemical structure is less clearly associated with the final adverse outcome.

Use of read across to assess mutagenicity in a Regulatory context – the EU REACH regulation

Read-across of hazard data between structurally similar analogues and within categories plays a critical role within the EU REACH regulation as it offers a way to identify potential hazards of the substance subject to this regulation while reducing the need for generating new studies on every substance. However, due to the importance of performing a reliable and robust hazard assessment to demonstrate a chemical substance can be used safely, the use of read-across must be supported with a robust scientific justification. In the absence of such a justification there is too much uncertainty associated with the hazard characterization.

A substantial amount of guidance and published papers are available on the subject of read-across to guide a practitioner through the process of grouping substances and formulating a robust justification [3, 4, 5, 6, 7] and on the side of the regulatory agency (the European Chemicals Agency), a guide for assessors has been developed, the Read-across Assessment Framework (RAAF) [8]. This guide was developed to support the assessment of read-across justifications, promoting a consistent approach and resulting in a transparent judgement on their adequacy and acceptability. In situations where the read-across justification is deemed to be insufficient, it also provides a clear basis for whether and how it can be improved.

At a high level, the RAAF is set up to take an assessor through a read-across justification such that the critical elements are assessed and given a score, allowing the assessor to conclude whether the overall justification is OK and, if not, what are the aspects of the justification which are insufficient or scientifically invalid. To accomplish this, the RAAF puts read-across justifications into six different scenarios depending on whether a category or analogue approach is used, what is the main rationale/hypothesis underpinning the read-across justification and is the prediction for a similar versus difference potency. For example, read-across within a category of structurally similar but different compounds having the same kind of effect and with a similar potency. Within each scenario several assessment elements (AEs) have been defined. These AEs represent critical aspects of a read-across justification that must be addressed and supporting information provided. For each AE a series of questions have been defined to prompt the assessor to look whether the necessary supporting information has been provided, or whether a concern has been addressed. For example, where metabolism is an important aspect of the read-across justification, one of the AEs guides the assessor through the determination of whether information on metabolism has been provided and is sufficient. These AEs are scored by the assessor with a score of 1-5; a score of 1 or 2 is given where assessor considers the information provided to be invalid or insufficient. A score of 3-5 indicates the information provided is accepted but with varying degrees of confidence/certainty – 5 indicating that the AE is accepted with low uncertainty. The culmination of the assessment is a series of scores for each AE and if any of these is 1 or 2 then the read-across justification is rejected. The AEs where scores of 1 or 2 are given form the basis of feedback to the registrant on whether their approach is considered scientifically valid and whether it is deemed possible to improve the justification by providing more information.

One critical aspect of the use of read-across within the context of REACH is that a read-across justification should be specific to each endpoint where read-across is used. For example, if read-across is used to address genotoxicity, repeated dose toxicity, environmental fate, etc. then for each endpoint there should be a specific justification provided to justify the use of read-across. These endpoint specific justifications can have common elements, but it should be specifically stated how structural similarities between analogues or category members allows the prediction of the specific endpoint in question. This has not always been clearly understood in the past and led to situations where read-across was accepted for one endpoint but not another [9].

The RAAF was originally intended to be a guide for the European Chemical Agency (ECHA) to use as part of dossier evaluations, however it is also a useful resource for practitioners preparing read-across justifications [10]. By identifying the most appropriate read-across scenario, the relevant AEs can be identified, and a practitioner can then ensure enough information is available to support them. Several published case studies are available where this approach has been employed along with attempts to characterize uncertainty associated with the use of read-across [11, 12, 13]. While making use of the RAAF to prepare a read-across justification should result in it being more robust, it does not guarantee acceptance of the approach by ECHA. There is still room for expert judgement with respect to the interpretation of information provided and the uncertainty associated with the approach.

2. Kinetics and PBPK modelling as a component for read across

The need for ADME/TK data in read-across

Read-across, the process by which information from known (source) chemical(s) is used to infer the activity of unknown (target) chemical(s), is important for chemical safety assessment across multiple industrial sectors (food, drugs, pesticides, household products, cosmetics etc.) as it provides a means to reduce animal testing, supporting the 3Rs philosophy as presented by Judith Madden and Alicia Paini. Whether it is to be used for internal decision-making, research and development or in regulatory submissions, determines the extent of justification required or level of uncertainty that is acceptable for the prediction, with regulatory applications being the most demanding.

In predicting true potential to elicit a biological response it is not only the intrinsic activity of the chemical that is relevant, but also its ability to reach the site of action in sufficient concentration i.e. the internal exposure of the chemical needs to be characterised. Hence, full justification of a read-across prediction requires consideration of the absorption, distribution, metabolism and elimination (ADME) properties of source and target chemicals, and ideally their time-dependent (toxicokinetic (TK)) profile. In order to assist read-across practioners to provide adequately justified predictions, guidance has been published on how these should be reported and documented, including how to incorporate relevant ADME/TK information. The Read-Across Assessment Framework from ECHA [2] defines a specific scenario for read-across, on the basis of conversion to a common metabolite. The OECD guidance on grouping of chemicals stipulates credibility of read-across is enhanced where similarity in ADME can be demonstrated and proposes that kinetic data may be used to demonstrate that source and target chemicals are dealt with by the body in a similar (or predictable) manner [26]. The document also provides an example data matrix for organising information used in the prediction. Similarly, the strategy for structuring and reporting a read-across prediction published by Schultz et al. provides template tables for inclusion of ADME data that may support a read across prediction [24].

In building a category, or identifying individual analogues for read-across, chemicals must be selected based on appropriate similarity metrics, with any dissimilarities identified. Whilst the argument centers on the similarity driving the read-across prediction, any dissimilarities that may moderate, or obviate, activity must be considered. It is a truism that no chemical can be absolutely similar to another – only similar with respect to a given property. Therefore, there are many properties that may be used to compare chemicals: structural similarity, biological similarity, physico-chemical properties, mechanism of action, metabolite formation etc. To have confidence in the read-across prediction, it is essential either to select only those analogues that have similar (or predictably differing) ADME/TK properties, or to moderate the prediction with any known or predicted ADME/TK data (for the source or target chemicals). Differences in testing scenario may require additional modifications e.g. extrapolation between different dose levels (low to high dose; single to multiple dose) exposure routes (oral, dermal or inhalational) and inter-species differences. Increasingly, there is interest in predicting intra-species differences as, for example, the response seen in a healthy adult may be significantly different to that seen in a neonate or aging person/animal (*vide infra*). Identifying the most sensitive individuals in a population is also important, as chemical safety assessment must establish safe levels for normal, population-wide, use.

Sources of ADME/TK data

Knowledge of ADME/TK parameters for source and target, notably extent of tissue or plasma protein binding, metabolic profile or elimination rate/half-life, is key to a well-justified read-across argument. With advances in experimental methods to generate data and an expansion in resources for storage (often in the public domain) there is an ever-growing knowledgebase for ADME/TK data. General collations such as PubChem (https://pubchem.ncbi.nlm.nih.gov/) with extensive data on 96 million compounds or ChemSpider (http://www.chemspider.com/) with data on 74 million compounds now provide vast amounts of information, rapidly searchable for target chemicals or potential source chemicals, identified using integral routines to identify structurally similar chemicals. Another major resource is the Computational Chemistry Dashboard (https://comptox.epa.gov/dashboard) that contains extensive *in vitro* assay data for 875, 000 compounds (at time of writing). With respect to collating ADME data specifically, Przybylak et al (2018) [22] compiled 140 ADME-related datasets. To assist with predicting values, Patel et al (2018) [19] reviewed over 80 published quantitative structure-activity relationship (QSAR) models for ADME-related properties. A subsequent paper by Madden et al (2019) [17] includes over 200 resources for obtaining or predicting ADME/TK properties. A more comprehensive review of over 950 *in silico* toxicology data resources to support read-across and (Q)SAR has been published by Pawar et al (2019) [20]. Metabolism data is arguably one of the most important factors to consider and there are multiple packages available to predict metabolites e.g. Meteor Nexus (Lhasa limited) and the OECD QSAR Toolbox (https://qsartoolbox.org) which has modules for predicting and/or retrieving metabolites in skin or liver, in rat or human. Toxtree (http://toxtree.sourceforge.net/) also possesses a module to predict potential sites of CYP 450 metabolism and putative metabolites. Although several packages exist to predict possible metabolites, a shortfall of these is their inability to predict the rate and extent of production of individual metabolites. If a read–across argument is based on metabolic similarity, generally this will be reliant upon inclusion of experimental data.

Examples of incorporating ADME/TK data in read-across

Case studies are useful to demonstrate the utility of new methods in resolving contemporary problems. In order for read-across to become a more accepted methodology, particularly with respect to regulatory submissions, example case studies are being developed and reviewed at the OECD level. One example of a read-across case study incorporating TK data is for sub-chronic repeated-dose toxicity of simple aryl alcohol alkyl carboxylic esters [27]. In this example three subcategories of aryl alkanoates were investigated: benzyl alkanoates, 2-phenylethyl alkanoates and 3-phenyl propyl alkanoates. The read-across argumentation for the category members (for chain lengths in the category from C2-C12) was that the ADME processes were well characterised i.e. that the alkanoates were metabolised to the corresponding alcohol which drives the toxicity (non-specific interactions with cell membranes; non-polar narcosis) and to carboxylic acids with little local and no systemic toxicity. The argument for read-across between the esters and the common metabolites, based on both toxicodynamic (TD) and toxicokinetic (TK) similarity, was supported with *in silico, in vitro* and *in vivo* metabolic data with the available *in vivo* data reducing uncertainty. An OECD report (OECD, 2018b), reviewing this and other case studies, concluded that using additional elements for the read-across i.e. going beyond structurally similarity and including TK and other data types strengthened the hypothesis. One caveat reported, however, was concern over the quality of the experimental data. This is an important issue for all read-across cases where a limited amount of *in vivo* data may have significant weighting on the overall argument. The report [28] proposed three areas for further development. These related to guidance on generating further *in vitro* data to support the argument; guidance for evaluating data reliability (for both TK and TD data); and guidance on addressing uncertainty. This third recommendation was the subject of a subsequent paper by Schultz et al (2019) [25] in which six case studies for repeated-dose toxicity were reviewed to identify, evaluate and assess the nature of uncertainty associated with the read-across arguments contained therein. From this analysis a series of 30 questions were proposed to help address the different types of uncertainty identified. The authors concluded that “similarity in toxicokinetics, especially in ADME properties, is seen as crucial in assessing uncertainty. Metabolism is often seen as the most contentious of the toxicokinetic similarity justification”. Two of the 30 questions addressed this issue specifically: (i) Is there sufficient ADME information provided to establish toxicokinetic similarity for the derivatives used in the read-across? (ii) Are any dissimilarities in ADME properties (and, as appropriate, metabolism / degradation) toxicologically relevant?

Other initiatives, aimed at increasing awareness of the role of toxicokinetic information in read-across, include the European Partnership for Alternative Approaches to Animal Testing (EPAA) Partners' Forum on Toxicokinetics and Read-Across. Lack of TK data was identified as an impediment to read-across and the authors proposed generating such data using *in vitro* and *in silico* tools. The report summarised the strategies for inclusion of TK data that are being employed by research organisations and industries internationally with specific examples of working practices being reported by European Commission Joint Research Centre (JRC), OECD and representatives from the cosmetics, fragrance, agrochemical, chemical sectors as well organisations representing the food and pharmaceutical industries. Overall recommendations from the forum included collating available tools for TK, providing guidance on the TK parameters necessary to support read-across and further use of case studies. It was recognised that physiologically-based kinetic (PBK) modelling was used more in some areas than others. In response to this report, the paper of Madden et al (2019) [17] provides an extensive list of tools that are available to predict TK properties and that may be used to support development and evaluation of PBK models.

Using data from physiologically-based kinetic (PBK) models in read-across

In a physiologically-based kinetic model, the body is divided into a series of compartments connected by blood flow. Effectively these models require a series of differential equations to be solved for individual organs relating to the concentrations entering and leaving each organ over time, taking into consideration factors such as organ volume, blood flow and the potential for excretion, metabolism and storage. The models can be adapted to different individuals within a population enabling physiological and anatomical differences to be taken into account. For example, metabolic difference are often responsible for observed differences in activity within a population as metabolic capacity in neonates, or those suffering age-related liver degeneration, are different to healthy adults. Other factors such as differences in gastro-intestinal transit, secretory processes, membrane permeability, plasma protein binding, total body water, glomerular filtration, renal tubular secretion/reabsorption also lead to age-related differences in the population [14]. In addition to the anatomical and physiological information, chemical specific information (e.g. physico-chemical properties) are also included, these are adapted for different chemicals. PBK models enable more accurate derivation of the concentration of a chemical within individual tissues over time for different members of the population. The models can be readily adapted for different species, routes of exposure and different dosing scenarios. Whilst incorporation of general TK parameters are important for supporting read-across, using data from PBK models provides a more holistic view of the internal exposure and hence potential effect, enabling effects to be correlated with dose at the target site rather than an association with an arbitrary external dose.

One drawback of the models is that they are highly data hungry, requiring optimisation of multiple factors in order to accurately reproduce concentrations within blood or other tissues. Leveraging information from existing models to make predictions for new chemicals would therefore be beneficial. Lu et al (2016) [16] developed a Knowledgebase of existing PBK models for 307 unique chemicals, suggesting these could provide templates from which PBK models for new chemicals could be derived. The information on time-concentration comparisons between target and source chemicals could be used to support a read-across hypothesis. As for all read-across argumentation, the selection of the most similar analogue(s) to use as a template is critical to the case. In the analysis of Lu et al physico-chemical properties were used to identify the most similar chemicals. However, it has been demonstrated, that using different similarity metrics will result in different chemicals being selected as “most similar”. As a proof-of-principle, five chemicals were investigated here. One chemical was selected at random from each of three databases representing drugs (www.drugbank.ca), cosmetic-related ingredients (https://cosmbosdb.eu) and food additives (http://foodb.ca) in addition to the two case study chemicals investigated by Lu et al. The five chemicals were: indinavir (drug); ethylparaben (anti-fungal preservative); allyl heptanoate (fruit flavoring); ethylbenzene and gefitinib - the case study chemicals from Lu et al. Nine similarity metrics (Morgan, feat Morgan, Torsion, Avalon, Layered, AtomPair, RDKit, MACCS and Pattern) from RDKit were used to find the most similar chemicals to these five target chemicals (considering only those chemicals for which an existing PBK model is available in the Lu et al Knowledgebase) [16].

As anticipated the use of different similarity metrics resulted in different chemicals being selected as most similar. Interestingly the chemical selected by Lu et al as being most similar to gefitinib, according to physico-chemical properties, never appeared in the top three most similar according to the nine RDKit similarity metrics. Typically, the same chemicals were identified as being most similar by three or four of the nine metrics, whereas the other metrics selected different chemicals. This is important as it demonstrates a fundamental issue in read-across i.e. the use of different similarity metrics to identify “similar” analogue(s) for read-across will result in different analogue(s) being selected. Hence, expert judgement and additional justification are required to ensure the most appropriate analogue(s) are chosen. This issue has already been widely reported [18] and different criteria for analogue selection have previously been proposed [29]. At the OECD level, in an international effort to promote the regulatory use of PBK models based on non-animal data [23] studies have been undertaken to evaluate PBK model predictions, where data from appropriate analogues have been used to assist model development. In effect, an existing PBK model is used to provide a PBK model template that can then be adapted using (physico-chemical) information for an alternative chemical of interest. The aim is to develop a strategy to assist chemical safety assessment where *in vivo* data are lacking; this work is on-going.

Future work on analogue selection will need to consider the most appropriate similarity metrics to use, these may include measures of chemical similarity, biological similarity or a combination of the two. Novel similarity metrics are discussed further below.

3. Structural similarities and molecular design

Background

Read-across (RA) is a technique for inferring endpoint information for one substance (target substance), by using data from the same endpoint from another substance or substances, (source substance), that has similar properties than the target substance. “Endpoints” can have different meanings depending on the context. In the context of the REACH information requirements, endpoints are described either as a property itself (e.g. skin irritation, long-term toxicity to sediment organisms) and/or as a type of study (e.g. carcinogenicity study, fish early life stage test) [14]. A key assumption is that molecules with high structural similarity produce similar toxic effects [15] and under REACH, any read-across approach must be based on structural similarity between the source and target substances [14].

Introduction

Computational methods as presented by Miriam Mathea and Andrew David Palmer provide chemical similarity measures are important for several fields of computational toxicology, because they can be used to predict the molecular properties of structurally close compounds. One very important application of chemical similarity measurements is the read-across approach [32]. Recent large benchmarking studies, such as the Tox21 challenge from the NIH, have made computer-readable assay data for ~10K compounds available and quantified a signification improvement in the quality of predictive models [20]. Almost all computational approaches require a step where a molecule is translated from a chemical structure into a linear digital representation, called a molecular fingerprint, which contains the information used to build a model. Therefore, any model with a high predictive value must be based on a fingerprint which captures the toxicological information. Structural similarity measures evaluate how alike pairs of molecules are based on their fingerprints. These similarity measures are important for computational toxicology, because they can be used to calculate which compounds are ‘structurally close’ and therefore likely to share toxicological properties, which is the essence of the read-across approach [61].

There are several approaches and algorithms available for calculating chemical similarity [16],[17]. They can be divided in different combinations of available molecular descriptors, distance/similarity measures and weighting schemata for specific endpoints [61] but it is not obvious which measures of structural similarity correlate with toxicological similarity. Hence, it is not straightforward for the user to select a method for his specific endpoint/s. It would be advantageous to have a molecular fingerprint that “adapts itself” to the data (endpoint/s) and provides good ability to distinguish between toxicological similar and not similar molecules.

In this study we will benchmark a Neural Representation method [19] with state of the art 2D fingerprints and physicochemical descriptors regarding their applicability for read across [17],[18].

In general a molecular fingerprint consists of a fixed length bit string in which the presence or absence of an attribute (i.e. substructure) is encoded by a hashing algorithm. In comparison to the applied physicochemical descriptor which is based on a combination of 117 physicochemical properties (Rdkit) [22]. We compare these two chemical descriptors with a Neural Representation, that learns the molecular representation, regarding their ability to identify relevant similar compounds for a given dataset. We use a k-nearest neighbor classification method to evaluate the discriminative power of the different descriptors. This method searches the closest molecules to the target molecules in the dataset and then compares their major vote over the label of the endpoints with the label of the target compound. In this example label of endpoint means “active” or “inactive” on a specific endpoint/target.

Dataset and Method

The dataset we used in this study was provided by the U.S. department of Health and Human services within the Tox21 challenge [20]. It includes the assay measurements for 12 different endpoints. In order to compare different techniques a training and a test set was provided to the participants. See [20] for details.

Physicochemical properties of the molecules were calculated with the 117 descriptors of RDKit [22], an open-source cheminformatics tool. These define type and number of atoms, bonds and rings present in the molecule, polarity and solubility, among other molecular properties. The topology of molecules was described by 1024 bit Morgan Fingerprints (ECFP) [23], 2D fingerprints encoding circular atom neighborhood in a hashed bit string. Thus, each bit of the fingerprint represents the presence or absence of a substructure.

Furthermore, a graph based neural networks were employed to define a learned molecular representation. This is constructed by operating on the graph structure of the molecule, i.e. it is an own feature representation learned directly from the data. The model we used has two distinctive features. The first one is it operates over a hybrid representation that combines convolutions and molecular descriptors. This design offers it flexibility in learning an endpoint specific encoding, while providing a strong prior with the help of fixed descriptors. The second one is it learns to construct molecular encodings by using convolutions centered on bonds instead of atoms [19].

In order to evaluate the performance of these three different molecular descriptor types, the k-nearest neighbor classification was used. The parameter k defines the number of neighbors the algorithm should select und was varied from 1 to 50 and Euclidean distance was used to measure the distance between pairs of molecular descriptors. For each molecule in the test set: the k nearest neighbors from the training set partition were obtained and either the major vote, weighted average or maximum over the neighbors' labels was computed.

To compare the different descriptor types the accuracy (correct classification rate), sensitivity (true positive rate) and specificity (true negative rate) were computed. The correct classification rate is the percentage of molecules which class label was correctly assigned to ("active" or "inactive"). The true positive rate describes in this example the number of active molecules that were correctly classified as active molecules and the true negative rate vice versa. Additionally, analysis using 2D Uniform Manifold Approximation and Projection (UMAP)[62] was performed and colored by the measured activity in the assay. Plotting the UMAP embedding is designed to provide a good visual overview how well the molecular descriptor might be able to distinguish between the labels as well as the coverage of the chemical space (Figure 1).

Results and Discussion

In principal, molecular similar molecules should cluster together when applying a principal component analysis based on a descriptor. But the quality of the clustering differs dependent on the type and the accuracy of the molecular descriptor encoding the molecules. Hence, the more detailed the molecular representation is the better the separation of the molecules when plotting the first two principal components. Following the principal of read-across that molecular similar molecules should lead to similar toxic effects, the best molecular descriptor should be the one that best clusters molecules together that have the same toxic effect, so in this case are active or inactive on a specific target. The analysis will be shown exemplarily for the estrogen nuclear receptor alpha [homo sapiens]. The results are shown in Figure 1. The graphic on the left-hand side represents the first two principal components based on the Morgan Fingerprint. The violet dots are the molecules that are active on the estrogen receptor and the red dots the molecules that are inactive. The graphic on the right-hand side shows the first two principal components based on the graph convolutional fingerprints. The active molecules in the plot of the graph convolutional fingerprints cluster closer together than in the plot of the Morgan fingerprint. Hence the graph convolutional fingerprint can better distinguish between the active and inactive molecules.

Figure 1 near here

Since the UMAP analysis only gives a visual impression of the performance of the molecular descriptors, the k-nearest neighbor analysis was applied to get more detailed insights about their characteristics.

We simulate a read-across situation for estrogen nuclear receptor activity using k-nearest neighbors analysis, where we infer the class of a test molecule based on the activity of similar molecules (where similarity depends on both the molecular representation and the distance measure). As we know the true class of the test molecules we can evaluate the most appropriate performance of all combinations. By varying the number of neighbors (k) we increase the number of molecules that must be considered for evaluation. The Tox21 dataset of biological activity contains far more inactive compounds than actives, and so in an imbalanced dataset. Hence specificity is important, as the statistical likelihood of picking an inactive compound is high. We can see from Figure 2 that for the weighted distance measure (see Figure 3 for all combinations) that the neural fingerprint consistently equals or outperforms the other two descriptors in terms of accuracy, sensitivity and specificity, regardless of how many neighbors are considered for analysis.

Figure 2 near here

Figure 3 near here

Conclusion

Defining molecular similarity in a toxicological context remains a challenging task and this pilot study illustrates that choice of fingerprinting approach is a critical aspect. The most appropriate molecular descriptor should capture sub-structural element that enable differentiation between active and inactive compounds. Our analysis shows that describing the latent molecular space using Neural Representations increases the read-across success rate. We attribute this to the fingerprints being learned from the underlying dataset and so producing an endpoint specific dense representation. Hence, they are customized for the specific read-across questions and can better distinguish between the active and inactive molecules. The trade-off is that the toxicological fingerprint is dependent on both the molecular and toxicological spaces whereas 2D fingerprints and physicochemical descriptors are dataset independent. As the read-across exercise takes place in the context of molecules with known toxicological properties it should be possible though collaborative efforts to establish a broad toxicological space which can be expanded as needed.

4. The use of metabolomics to improve the quality of read across

Read-across, as described in the ECHA read-across assessment framework (RAAF), is based on structural similarity of the target and source substance. Structural similarity does not necessarily lead to similar toxicological effects. A prominent example indicating differential toxicological effects despite high structural similarity, are 2-Acetylaminofluorene (2-AAF) and 4-Acetylaminofluorene (4-AAF). 2-AAF is a strong liver enzyme inducer thereby inducing tumor formation in the liver, whereas 4-AFF is only a mild liver enzyme inducer and non-carcinogenic (Table 1) (ToxNet 2-AAF, ToxNet 4-AAF; https://toxnet.nlm.nih.gov/).

Table 1 near here

Since already small changes in the structure of a molecule can lead to differential biological effects, an inherited insecurity is present for every read-across approach. Therefore, every read-across submitted to ECHA needs to be supported by further data proving that also on a biological level the substances behave similarly. Nevertheless, most applicant fail to provide this kind of supportive data in a sufficient manner leading to a high dismissal rate of submitted read-across cases [42].

Saskia Sperber introduced Metabolomics as a data-rich ‘omics-technology, describing analogous to other ‘omics-technologies, such as transcriptomics or proteomics, the analysis of the complete set of small-molecule metabolites (metabolome) that specific cellular processes leave behind within a biological sample [39]. The metabolome has the advantage that it is placed downstream of the transcriptome and proteome, integrating the regulatory steps of upstream levels of organization, thereby reflecting the actual phenotype better than other ‘omics techniques [40]. Furthermore, the metabolites can be analyzed in a relatively non-invasive manner in different body fluids, such as, urine or blood, giving information about numerous organs at the same time as well as enabling time-resolved analysis in the same animal to confirm toxicological manifestation or adaption processes on a single animal basis [41]. Metabolomics, due to the above-mentioned reasons, is a suitable technique to add valuable biological data as supporting information for structure based read-across approaches.

BASF SE has already started in 2004 to build up a database, called MetaMap® Tox, with metabolic profiles resulting from highly standardized 28-day toxicity studies with now more than 1000 chemicals, agrochemicals and pharmaceuticals. The study protocol has been described by van Ravenzwaay et al. and is most similar to an OECD 407 guideline study to enable integration of the procedure into these without need for further animal testing [42]. Thereby, the evaluation process for a new test compound is conducted in a three step process: 1) The biochemical analysis of the changed metabolites 2) A comparison of the metabolic profile of interest with all other available metabolite profiles available in the database (Profile comparison) and 3) a comparison with end-point specific metabolite patterns created from well characterized reference compounds, as described in detail by Sperber et al. [43]. With the obtained data from this analysis target organs and toxicological modes of action can be predicted. However, the biggest advantage of the MetaMap®Tox database for read-across are the more than 1000 metabolic profiles for different treatments available in the database and the direct comparison between these. This algorithm enables us to identify the most similar treatment in the database regarding their effects on the plasma metabolome.

As a proof-of-concept study van Ravenzwaay et al. published a hypothetical read-across scenario with phenoxy herbicides in 2016 [44]. In this case study Mecoprop (MCPP) was pretended to be the target substance with missing data for a 90-day study (OECD 408 guideline study), whereas these data were available for Dichlorprop (2,4-DP) and MCPA, the designated source substances. For all three treatments toxicological data from a 28-day study (OECD 407 guideline study) as well as metabolome data was available. Van Ravenzwaay et al. showed that all three treatments induced very similar metabolic changes and could correctly identify the liver by peroxisome proliferation and the kidney by inhibition of the organic anion transporter (OAT1) as the target organs. Nevertheless, the above described profile comparison, a principal component analysis (PCA) as well as a direct comparison of the individual metabolite changes all confirmed 2,4-DP to be the best source substance for a read-across approach, which was well confirmed by the actual findings of the 90-day study for MCPP.

In 2019 Sperber et al. published a follow up study with 2-aminoethanol (MEA) (source substance) and 3-aminopropanol (3AP) (target substance) delivering all relevant data for a real read-across approach as it could be submitted under REACh [43]. It could be shown, that 2-aminoethanol, from all available treatments in MetaMap® Tox is the most similar one and that other structurally similar chemicals, such as Diethanolamine and Triethanolamine, induced significantly different effects in the metabolome of treated rats. Interestingly Sperber et al. also took into account the quantitative aspect and could show that even though the induced effects were overall similar at comparable dose levels stronger effects were observed for the source substance.

Metabolomics can thereby also help to proof our hypothesis on which source substance might be the most suitable for a respective target substance. Based on structural similarity, BASF Substance 2 was hypothesized to represent the best read-across option for the target substance BASF Substance 1. However, the available metabolome data told a different story and only minimal commonly induced metabolite changes could be detected for both treatments (Figure 4). BASF Substance 3 and 4, which also share a highly similar chemical structure with BASF Substance 1 were found in the highest ranks of the profile comparison sharing a large fraction of commonly changed metabolites. These findings show metabolomics to be able to confirm biological similarity of different treatments and how it can be used to support decision making on the best source substance for a potential read-across.

Figure 4 near here

As well as indicating biological similarity of different treatments, metabolomics can also be used to identify differences between toxicological modes of action as shown by Langsch et al. [8]. An *in vitro* study by Campioli et al. suggested Hexamoll® DINCH, an important non-phthalate plasticizer, to lead to similar perturbances in fat storing as Diethylhexylphthalate (DEHP) ultimately causing obesity [46]. Langsch et al. combined classical *in vivo* toxicity studies with *in vivo* metabolome data to disprove this hypothesis. Metabolomics could clearly show Hexamoll® DINCH to induce significantly different metabolic changes compared to DEHP, clustering even close to control animals.

Overall, metabolomics was shown to be a valuable tool to add supportive biological information to structure-based read-across approaches, especially in the context of BASF's MetaMap® Tox database including metabolite profiles for more than 1000 different treatments. Biological comparison of treatment effects can thereby significantly reduce the inherited insecurity of structure-based read-across approaches.

5. Big Data In silico tools for read across

Read-across, i.e. the inference of toxicological properties from similar chemicals, has become a key data-gap-filling approach over the last two decades, especially in the context of the European REACH legislation. In contrast, *in silico* and *in vitro* approaches have not been used to a large extent. Originally, read-across was a pragmatic, manual process, which depended strongly on the expert involved varying strongly in data, which happened to be available, and in execution and documentation. Not astonishingly, the acceptance rates in the REACH context were rather low [5]

Thomas Hartung and colleaguesof the Center for Alternatives to Animal Testing (CAAT) at Johns Hopkins University, became active in this field in 2013, starting with assembling an expert group from industry, regulators and academia. A white paper [4] was elaborated to map needs and opportunities in the area. It became clear that a key problem was that guidance how to do read-across properly was missing and expertise was with few individuals. Many of these were recruited for a taskforce to develop to develop Good Read-Across Practice (GRAP) [5]. This guidance was presented and endorsed in two stakeholder fora in Brussels and Washington with more than 400 participants. GRAS complements ECHA's RAAF, published in September 2015. It became clear that an enormous potential lies to go beyond chemical structural similarity and include also biological similarity. However, consensus was reached that no general guidance could be developed for biological read-across. An accompanying document [60] was therefore created to illustrate with promising examples, how biological read-across could be set into practice.

The process identified as a main hurdle for high-quality the availability and accessibility of data on similar chemicals. Only few large companies can rely on sufficient in-house data. Most data on chemicals in the past have been proprietary and those publicly available are biased toward pharmaceuticals, few intensively studied ones with interesting (complex) mechanisms of action and especially toxic substances. We rarely see publications that a substance has no effect, also given the impossibility to exclude action ("absence of evidence, is no evidence of absence"). The very legislation, REACH, which fueled the interest and use of read-across also included as a first world-wide that registration comes with making at least summary data on the safety assessments publicly available. However, at the time, the respective dashboard of the European Chemical Agency (ECHA) provided these as little structured, not machine-readable information, accessible only one chemical at the time in a manual process. The download of these data was explicitly prohibited. We felt that this treasure trove of chemical safety data needs to be available for computational toxicology and read-across. We downloaded the public data in December 2014 in a way that data flow on the website was not impaired. Using natural language processing, data from 800,000 toxicological studies on almost 10,000 chemicals were obtained in a machine-readable database [51] The European Commission and ECHA were informed about the upcoming publication of four publications analyzing these data [51]. This resulted initially in some irritation by ECHA [47], which resulted in their request not to publish the database as originally intended. We complied and the papers were published without the database. In a number of high-profile press coverages, the value of the database was praised, e.g. see Rabesandratana (2016) [56].

In a meeting two months later at ECHA, common ground was found. We agreed that availability of such data is in the best interest of chemical safety sciences and animal welfare. However, following ECHA's concerns, we have not published the database ourselves but make it available to interested parties in "collaborations". A planned joint press activity around ECHA making the data officially available through us faltered as University legal counsel was not willing to accept liability for this indemnifying ECHA. ECHA did publish therefore about a year later a database on an enlarged number of 15,000 chemicals, though redacted with respect to some information on the website. Most importantly, ECHA clarified separately that these data can be used in aggregate manner to generate information for registration purposes: Chemical Watch 5 July 2017 - "*ECHA gives clarity on IP issues for QSAR predictions … A registrant would need permission to use protected data to read-across from a single substance to the target substance, … But they would not need this to make a Qsar prediction.*"

This opened up for the development of predictive tools using these data. Seeing the enormous value of such registration data, through our European policy program, we worked with members of the European Parliament toward broader release of agencies of registration data. The Environment Committee of the European Parliament adopted this position unanimously in fall 2017 and held a hearing with the heads of the relevant agencies in early 2018. The process has not been completed but following some positive responses a pilot project is currently in preparation.

Already the initial publications [51] illustrated the potential for an automated read-across. Different to a traditional read-across, where a formula is derived from chemical descriptors to make a prediction, which often only holds for relatively small parts of the chemical universe [48], the new approach is looking only into the similar chemicals to derive a prediction, similar to a read-across. We termed this a RASAR (read-across-based structure/activity relationship). Teaming up with Underwriters Laboratories (UL), a safety standard and testing organization and creating the spin-off ToxTrack LLC, this approach was further developed. The process is documented in a number of articles [49, 52, 53]. Informed choices had to be made as to how to expand the database, express chemical similarity, which requires chemical finger printing, the integration of (different types of) information on neighboring chemicals, the machine learning approach and the expression of results. Briefly, we combined several reliable data sources (PubChem, ECHA, ICE...). We used the most common Chemical Similarity (PubChem2d) with Tanimoto (Jaccard) metric; we tested 11 different metrics and remarkably Tanimoto gave best results as a stand-alone, but combinations of different metrics appear to have potentials for future improvements. We employed network features using proximity to positive and negative neighbors. Ultimately, Data Fusion was applied making use of other toxicity, biological and chemophysical endpoints not only the property to be predicted; in total 74 properties were included, which means that each chemical is characterized by a 222-dimensional vector (74 for the chemical itself and 74 for closest negative and positive neighbor, each). Machine Learning (logistic regression, random forest) gives probabilistic hazard estimates using Computing Clusters (Apache Spark pipeline) allowing massive scale computing. To illustrate the enormous computational effort, it took 180 core Amazon cloud server two days to calculate this similarity map for 10 million structures in our database [53] at about $5,000 in computing costs. More than 10 trillion one-to-one comparisons had to be made. The resulting map has chemicals, which are similar to each other close and those less similar distant to each other. Now, any chemical included in the 10 million structures or not can be placed into the map with about with "only" half a billion operations in less than a second.

To evaluate the quality of predictions, all chemicals with known classification were predicted in a five-fold cross-validation, which means that 190,000 predictions were made pretending we had no information on the substance and then comparing prediction and known result. This was done for all chemicals, even where no close or contradictory information on neighbors was available. A remarkable 87% correct results were obtained for nine common hazards.

In parallel, the original ECHA database was scrutinized for chemicals, for which multiple animal test results were available. 350-750 chemicals with repeat tests each were found for the six most commonly used toxicity tests. They were on average 81% reproducible, but this number is inflated by the non-toxic substances, which typically remain negative in re-tests. Only 69% reproducibility was found for toxic chemicals [53]. This apparent outperforming of the animal test with a computational method (81% vs 87% correct results) spurred a lot of media attention, e.g. Zainzinger (2018) [59] and van Noorden (2018) [57], though we have to admit that to some extent apples and organs are compared (not the same chemicals, 6 vs. 9 tests, animal test results vs. classifications). However, animal test reproducibility is a rather optimistic proxy for the reliability of test results anyway as all species differences are left aside in this analysis.

Further corroboration for the relevance of the approach comes from a type of sensitivity analysis provided in the publication: Analyzing which information contributed most to the prediction of the nine hazards, many of these made obvious sense. A key feature of the approach is, that a probability of the result can be calculated as well. The computer can analyze based on quality and quantity of results on the substance and its neighbors, how often the 190,000 predictions were correct or not. This allows to predict accuracy for any constellation of data. Already earlier we had seen for skin sensitization that the closer the next neighbor, the more likely the two substances were sharing the classification as toxic or non-toxic [49]. If we required 75% similarity, the prediction was 80% accurate, if we required 95% or more similarity, we obtained the same property in 92% of the cases. Defining the applicability of the approach based on the probability of an accurate result would boost the quality of predictions but on the expense of more and more substances, for which no prediction can be made. The other way around, the predictive power can obviously be improved adding more and more data, increasing the likelihood to have close chemicals with neighbors. Simulations show [53] that relatively small numbers of chemicals with data, talking hundreds to few thousands, allow to have similar chemicals for large parts of the chemical universe. The approach is allowing to work with very sparse datasets.

The potential for use for example in Green Toxicology [55] is obvious. The database also allows to derive improved thresholds of toxicological concern [58]. Since the publication, a number of extensions, improvements and validation efforts took place. Agencies such as FDA announced evaluation plans, but to our surprise we could not yet gain interest by International validation bodies to enter formal validation. The respective publications are in preparation. Altogether, the project represents an illustration of the potential of Big Data and Artificial Intelligence, which obviously goes far beyond safety sciences [37]. With a larger scientific community adopting and further developing the RASAR approach, this promises to be a major addition to the toolbox of safety sciences.

DISCUSSION AND CONCLUSION

The first goal of the workshop was to provide general information about the use of chemical grouping for read across purposes. The basic framework is laid out by ECHA's 2015 RAAF. Further guidance was provided by a taskforce which developed a guidance paper related to Good Read-Across Practice (GRAP) [6]. This guidance was presented and endorsed in two stakeholder fora in Brussels and Washington and complements ECHA's Read-Across Assessment Framework

Subsequently examples were given for classical read across for mutagenicity. Mutagenicity is usually related to reactivity (i.e. interaction of the chemical with the DNA) and as such has a higher likelihood of having a chemical structure-based similarity than for other toxicological end points such as systemic toxicity or reproduction toxicity. In the second presentation the toxicokinetic and metabolism aspects which need to be considered were highlighted. It is particularly noted this area that is often less well addressed, and its complexity underestimated. Therefore, insufficient ADME information may be a frequent cause for regulatory non-acceptance of proposed read across cases under REACH. In the third presentation the essential component of read across – chemical similarity – was further elucidated. Here computational methods that provide indications of chemical similarity were discussed. These methods calculate physicochemical properties of molecules defining type and number of atoms, bonds and rings present in the molecule, polarity and solubility, among other molecular properties. The analysis shows that graph convolutional fingerprints are useful for read-across. One advantage is that the fingerprints are learned from the underlying dataset and can be generated endpoint specific. Hence, they are customized for the specific read-across questions and can better distinguish between the active and inactive molecules. Thus, computational methods are available to assist and evaluate best selection of chemical for read across purposes based on physicochemical parameters.

As indicated above, the number of successful read-across cases in systemic toxicity is limited. Therefore, additional information – derived from new approach methods (NAMs) can potentially be helpful to substantiate read-across. One such a NAM is metabolomics. Appling this technology in the context of regulatory 28-day studies in rats provided additional biological information to prove or disprove similarity on a biological basis. Such information was considered as very useful in the ECHA NAM workshop. In this 2016 Helsinki workshop a number of key suggestions to further apply NAMs in a regulatory context were made. There is a need for standardisation of NAMs as well as a better understanding of their relevance through thorough analysis of their performance and definition of their applicability. Documentation and/or access to the underlying data and algorithms. In addition, reporting templates for NAMs are required to encourage their use. NAMs were demonstrated to provide pertinent information relating to mechanisms of action i.e. toxicodynamics; however, fewer examples of their use for toxicokinetics were available. Paini et al 2019, discuss the need for development and potential application of biokinetic models using information from 'Next Generation' alternatives. The regulatory use of NAMs is anticipated to increase in a variety of applications and to address a number of regulatory challenges, including supporting read-across, prioritisation and screening; however, they should be applied with a full understanding of their potential advantages (e.g. rapid screening, improved mechanistic understanding) and limitations. To increase uptake and acceptance amongst all stakeholders, case studies and capacity building are required. In the context of NAMs used for read-across, the main outcomes and conclusions of the workshop were:

(1) Data from NAMs were shown to support read-across as well as providing useful and usable information for screening and prioritisation.

(2) There is a need for standardisation of NAM approaches so that they may be made transferable and transparent.

(3) The intrinsic quality and coverage of NAM data have to be defined and addressed.

(4) There is a need to understand and characterize uncertainty from NAM data and how these will affect WoE.

(5) There is a need for further case studies to demonstrate the practical application of NAMs.

Finally, the last presentation, takes the concept of read across one step further, from chemical grouping based on preselected similarity (albeit with computational aid) to an automated process taking into account both chemical similarity as well as biological outcome for the end-point of interest. Different to a traditional read-across, where a formula is derived from chemical descriptors to make a prediction, which often only holds for relatively small parts of the chemical universe [48], the new approach is looking only into the similar chemicals to derive a prediction, similar to a read-across - termed RASAR (read-across-based structure/activity relationship).

References

[1] Dearfield KL; Gollapudi BB; Bemis JC; Benz RD; Douglas GR; Elespuru RK; Johnson GE; Kirkland DJ; LeBaron MJ; Li AP; Marchetti F; Pottenger LH; Rorije E; Tanir JY; Thybaud V; van Benthem J; Yauk CL; Zeiger E; Luijten MNext generation testing strategy for assessment of genomic damage: A conceptual framework and considerations. Environ Mol Mutagen. 2017, 06; 58(5):264-283.

[2] ECHA 2017. Guidance on Information Requirements and Chemical Safety Assessment, Chapter R.7a: Endpoint specific guidance, version 6.0. July 2017; ECHA-17-G-18-EN

[3] Patlewicz G; Ball N; Booth ED; Hulzebos E; Zvinavashe E; Hennes C.:Use of category approaches, read-across and (Q)SAR: general considerations. Regul Toxicol Pharmacol. 2013, Oct; 67(1):1-12.

[4] Patlewicz G; Ball N; Becker RA; Booth ED; Cronin MT; Kroese D; Steup D; van Ravenzwaay B; Hartung T:Read-across approaches--misconceptions, promises and challenges ahead. ALTEX. 2014; 31(4):387-96.

[5] Patlewicz G; Ball N; Boogaard PJ; Becker RA; Hubesch B. Building scientific confidence in the development and evaluation of read-across. Regul Toxicol Pharmacol. 2015, Jun; 72(1):117-33.

[6] Ball N; Cronin MT; Shen J; Blackburn K; Booth ED; Bouhifd M; Donley E; Egnash L; Hastings C; Juberg DR; Kleensang A; Kleinstreuer N; Kroese ED; Lee AC; Luechtefeld T; Maertens A; Marty S; Naciff JM; Palmer J; Pamies D; Penman M; Richarz AN; Russo DP; Stuard SB; Patlewicz G; van Ravenzwaay B; Wu S; Zhu H; Hartung T.: Toward Good Read-Across Practice (GRAP) guidance. ALTEX. 2016; 33(2):149-66.

[7] Blackburn K; Stuard SB. :A framework to facilitate consistent characterization of read across uncertainty. Regul Toxicol Pharmacol. 2014, Apr; 68(3):353-62.

[8] ECHA. 2015. Read-across Assessment Framework (RAAF). ECHA-15-R-07-EN

[9] Ball N; Bartels M; Budinsky R; Klapacz J; Hays S; Kirman C; Patlewicz G.:The challenge of using read-across within the EU REACH regulatory framework; how much uncertainty is too much? Dipropylene glycol methyl ether acetate, an exemplary case study. Regul Toxicol Pharmacol. 2014, Mar; 68(2):212-21.

[10] Schultz, T. W., Amcoff, P., Berggren, E., Gautier, F., Kalric, M., Knight, D. J., Mahony, C., Schwarz, M., White, A., Cronin, M. T. D. 2015. A strategy for structuring and reporting a read-across prediction of toxicity. Regul. Toxicol. Pharmacol. 72, 586-601.

[11] Gelbke HP; Ellis-Hutchings R; Müllerschön H; Murphy S; Pemberton M.: Toxicological assessment of lower alkyl methacrylate esters by a category approach. Regul Toxicol Pharmacol. 2018, Feb; 92:104-127

[12] Skare JA; Blackburn K; Wu S; Re TA; Duche D; Ringeissen S; Bjerke DL; Srinivasan V; Eisenmann C: Use of read-across and computer-based predictive analysis for the safety assessment of PEG cocamines. Regul Toxicol Pharmacol. 2015, Apr; 71(3):515-28. [Regulatory toxicology and pharmacology : RTP] [PubMed

[13] Schultz TW; Cronin MTD: Lessons learned from read-across case studies for repeated-dose toxicity. Regul Toxicol Pharmacol. 2017, Aug; 88:185-191.

[14] E. Fernandez, R. Perez, A. Hernandez, P. Tejada, M. Arteta, J.T. Ramos (2011) Factors and Mechanisms for Pharmacokinetic Differences between Pediatric Population and Adults, *Pharmaceutics*, 3(1): 53–72

[15] C. Laroche, M. Aggarwal, H. Bender, P. Benndorf, B. Birk, J. Crozier, G. Dal Negro, F. De Gaetano, C. Desaintes, I. Gardner, B. Hubesch, A. Irizar, D. John, V. Kumar, A. Lostia, I. Manou, M. Monshouwer, B.P. Müller, A. Paini, K. Reid, T. Rowan, M. Sachana, K. Schutte, C. Stirling, R. Taalman, L. van Aerts, R Weissenhorn, U.G. Sauer (2018) Finding synergies for 3Rs – Toxicokinetics and read-across: Report from an EPAA partners' Forum, *Regulatory Toxicology and Pharmacology,* 99, 5–21 https://doi.org/10.1016/j.yrtph.2018.08.006

[16] J. Lu, M-R Goldsmith, C.M. Grulke, D.T. Chang, R.D. Brooks, J.A. Leonard, M.B. Phillips, E.D. Hypes, M.J. Fair, R. Tornero-Velez, J. Johnson, C.C. Dary, Y.-M. Tan (2016) Developing a physiologically-based pharmacokinetic model knowledgebase in support of provisional model construction, PLoS Comput. Biol. 12 (2) e1004495.

[17] J.C. Madden, G. Pawar, M.T.D. Cronin, S.Webb, Y-M Tan, A. Paini (2019) In silico resources to assist in the development and evaluation of physiologically-based kinetic models, *Computational Toxicology,* 11, 33-49, https://doi.org/10.1016/j.comtox.2019.03.001

[18] C.L. Mellor, R.L. Marchese Robinson, R. Benigni, D. Ebbrell, S.J. Enoch, J.W. Firman, J.C. Madden, G. Pawar, C. Yang, M.T.D. Cronin (2019) Molecular fingerprint-derived similarity measures for toxicological readacross: Recommendations for optimal use, *Regulatory Toxicology and Pharmacology*, 101, 121-134

[19] M. Patel, M.L. Chilton, A. Sartini, L. Gibson, C. Barber, L. Covey-Crump, K.R. Przybylak, M.T.D. Cronin, J.C. Madden (2018) Assessment and Reproducibility of Quantitative Structure−Activity Relationship Models by the Nonexpert, *Journal of Chemical Information and Modelling*, 58, 673-682

[20] G. Pawar, J.C. Madden, D. Ebbrell, J.W. Firman, M.T.D. Cronin (2019) In Silico Toxicology Data Resources to Support Read-Across and (Q)SAR, *Frontiers in Pharmacology* (accepted)

[22] K.R. Przybylak, J.C. Madden, E. Covey-Crump, L. Gibson, C. Barber, M. Patel, M.T.D. Cronin (2018) Characterisation of data resources for in silico modelling: benchmark datasets for ADME Properties, *Expert Opinion on Drug Metabolism & Toxicology*, 14, 169-181, https://doi.org/10.1080/17425255.2017.1316449

[23] M. Sachana (2019) An international effort to promote the regulatory use of PBK models based on non-animal data, *Computational Toxicology,* 11, 23-24

[24] T.W. Schultz, P. Amcoff, E. Berggren, F. Gautier, M. Klaric, D.J. Knight, C. Mahony, M. Schwarz, A. White, M.T.D. Cronin (2015) A strategy for structuring and reporting a read-across prediction of toxicity. *Regulatory Toxicology and Pharmacology* 72, 586–601, http://dx.doi.org/10.1016/j.yrtph.2015.05.016

[25] T.W. Schultz, A-N. Richarz, M.T.D. Cronin (2019) Assessing uncertainty in read-across: Questions to evaluate toxicity predictions based on knowledge gained from case studies. *Computational Toxicology,* 9, 1-11, https://doi.org/10.1016/j.comtox.2018.10.003

[26] OECD (2014), Organisation for Economic Cooperation and Development (OECD), Guidance on Grouping of Chemicals, Second Edition, No. 194, Series on Testing & Assessment. ENV/JM/MONO(2014)4, OECD, Paris, 2014

[27] OECD (2018a), A case study on the use of integrated approaches for testing and assessment for sub-chronic repeated-dose toxicity of simple aryl alcohol alkyl carboxylic esters: read across, Series on Testing and Assessment No. 293

[28] OECD (2018b) Report on considerations from case studies on integrated approaches for testing and assessment (IATA), Third Review Cycle, Series on Testing and Assessment, No. 289

[29] S. Wu, K. Blackburn, J. Amburgey, J. Jaworska, T. Federle (2010) A f framework for using structural, reactivity, metabolic and physicochemical similarity to evaluate the suitability of analogs for SAR-based toxicological assessments, *Regulatory Toxicology and Pharmacology*, 56, 67-81

[30] Floris, M., Manganaro, A., Nicolotti, O., Medda, R., Mangiatordi, G. F., and Benfenati, E. (2014). A generalizable definition of chemical similarity for read-across. J. Cheminform. 6, 39. doi: 10.1186/s13321-014-0039-1.

[31] Zhu H, Bouhifd M, Kleinstreuer N, et al. Supporting read-across using biological data. ALTEX. 2016;33:167–182. http://dx.doi.org/10.14573/altex.1601252

[32] Nikolova N, Jaworska J. Approaches to measure chemical similarity - a review. QSAR Comb Sci. 2003;22:1006–1026. doi: 10.1002/qsar.200330831

[33] Bender A, Jenkins JL, Scheiber J, Sukuru SCK, Glick M, Davies JW. How similar are similarity searching methods? A principal component analysis of molecular descriptor space. J Chem Inf Model. 2009;49:108–119. doi: 10.1021/ci800249s.

[34] Willett P. Similarity searching using 2D structural fingerprints. Methods Mol Biol. 2011;672:133–158. doi: 10.1007/978-1-60761-839-3_5.

[35] Yang K., Swanson K., Jin W., Coley C., Eiden P., Gao H., Guzman-Perez A., Hopper T., Kelley B., Mathea M., Palmer A., Settels V., Jaakkola T., Jensen K., and Barzilay R. J Chem Inf Model 2019 59 (8), 3370-3388.doi: 10.1021/acs.jcim.9b00237

[36] https://tripod.nih.gov/tox21/challenge/

[37] Landrum, G. RDKit: Open-Source Cheminformatics; http://www.rdkit.org.

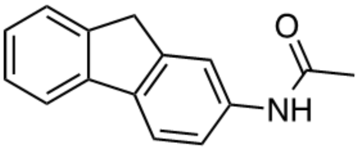
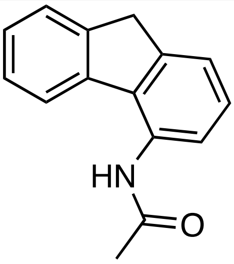
[38] Rogers D, Hahn M. Extended Connectivtiy-Fingerprints. J. Chem. Inf. Model. 2010;50(5):742-754. doi: 10.1021/ci100050t.

[39] S.G. Oliver, M.K. Winson, D.B. Kell, F. Baganz, Systematic functional analysis of the yeast genome, Trends Biotechnol, 16 (1998) 373-378.

[40] E. Fukusaki, Application of Metabolomics for High Resolution Phenotype Analysis, Mass Spectrom (Tokyo), 3 (2014) S0045.

[41] T. Cheng, X. Zhan, Pattern recognition for predictive, preventive, and personalized medicine in cancer, EPMA J, 8 (2017) 51-60.
[42] B. van Ravenzwaay, M. Herold, H. Kamp, M.D. Kapp, E. Fabian, R. Looser, G. Krennrich, W. Mellert, A. Prokoudine, V. Strauss, T. Walk, J. Wiemer, Metabolomics: a tool for early detection of toxicological effects and an opportunity for biology based grouping of chemicals-from QSAR to QBAR, Mutat Res, 746 (2012) 144-150.
[43] S. Sperber, M. Wahl, F. Berger, H. Kamp, O. Lemke, V. Starck, T. Walk, M. Spitzer, B.V. Ravenzwaay, Metabolomics as read-across tool: An example with 3-aminopropanol and 2-aminoethanol, Regul Toxicol Pharmacol, 108 (2019) 104442.
[43] B. van Ravenzwaay, S. Sperber, O. Lemke, E. Fabian, F. Faulhammer, H. Kamp, W. Mellert, V. Strauss, A. Strigun, E. Peter, M. Spitzer, T. Walk, Metabolomics as read-across tool: A case study with phenoxy herbicides, Regul Toxicol Pharmacol, 81 (2016) 288-304.
[45] A. Langsch, R.M. David, S. Schneider, S. Sperber, V. Haake, H. Kamp, E. Leibold, B.V. Ravenzwaay, R. Otter, Hexamoll((R)) DINCH: Lack of in vivo evidence for obesogenic properties, Toxicol Lett, 288 (2018) 99-110.
[46] E. Campioli, T.B. Duong, F. Deschamps, V. Papadopoulos, Cyclohexane-1,2-dicarboxylic acid diisononyl ester and metabolite effects on rat epididymal stromal vascular fraction differentiation of adipose tissue, Environ Res, 140 (2015) 145-156.
[47] Gilbert, N. (2016) Legal tussle delays launch of huge toxicity database. Nature News. http://www.nature.com/news/legal-tussle-delays-launch-of-huge-toxicity-database-1.19365 (Accessed 14 Dec 2016)
[48] Hartung T and Hoffmann S. Food for thought on…. in silico methods in toxicology. ALTEX 2009, 26:155-166.
[49] Hartung T. Making big sense from big data in toxicology by read-across. ALTEX, 2016, 33:83-93.
[50] Hartung T. Making big sense from big data. Frontiers in Big Data, 2018, Frontiers in Big Data 1:5. doi: 10.3389/fdata.2018.00005
[51] Luechtefeld T, Maertens A, Russo DP, Rovida C, Zhu H and Hartung T. Global analysis of publicly available safety data for 9,801 substances registered under REACH from 2008-2014. ALTEX 2016, 33, 95-109. http://doi.org/10.14573/altex.1510052.
[52] Luechtefeld T and Hartung T. Computational Approaches to Chemical Hazard Assessment. ALTEX 2017, 34:459-478.
[53] Luechtefeld T, Rowlands C and Hartung T. Big-data and machine learning to revamp computational toxicology and its use in risk assessment. Toxicological Research 2018, 7:732-744, doi:10.1039/C8TX00051D.
[54] Luechtefeld T, Marsh D, Rowlands C and Hartung T. Machine learning of toxicological big data enables read-across structure activity relationships (RASAR) outperforming animal test reproducibility. Toxicological Sciences, 2018, 165:198-212. doi: 10.1093/toxsci/kfy152.
[55] Maertens A and Hartung T. Green toxicology – know early about and avoid toxic product liabilities. Toxicol. Sci. 2018, 161:285–289. DOI: 10.1093/toxsci/kfx243.
[56] Rabesandratana, T. (2016). TOXICOLOGY. A crystal ball for chemical safety. Science 351, 651.
[57] van Noorden R. Software improves toxicity tests - machine learning trumps animal testing for many chemicals. Nature 2018, 559:163–163. http://doi.org/10.1038/d41586-018-05664-2
[58] van Ravenzwaay, B., Jiang, X., Luechtefeld, T., and Hartung, T. (2017). The Threshold of Toxicological Concern for prenatal developmental toxicity in rats and rabbits. Regulat. Pharmacol. Toxicol. 88, 157–172.
[59] Zainzinger V. Digital chemical test impresses - giant database shows promise for replacing animal studies. Science 2018, 361(6398), 117–117.
[60] Zhu H, Bouhifd M, Kleinstreuer N, Kroese ED, Liu Z, Luechtefeld T, Pamies D, Shen J, Strauss V, Wu S and Hartung T. Supporting read-across using biological data. ALTEX 2016, 33, 167-182. http://doi.org/10.14573/altex.1601252
[61] ECHA (2017), European Agency (ECHA) Read-Across Assessment Framework (RAAF). European Chemicals Agency, Helsinki, 2017.
[Becht E. et al Dimensionality reduction for visualizing single-cell data using UMAP. Nature Biotech, PY – 2018;37; 38 10.1038/nbt.4314.

Table 2: Structural and toxicological comparison of 2-acetylaminofluorene (2-AAF) and 4-acetylaminofluorene (4-AAF).

	2-AAF	**4-AAF**
STRUCTURE		
ADVERSE OUTCOMES	Strong liver enzyme inducer	Weak liver enzyme inducer
	Liver carcinogen	No liver carcinogen
	Immune suppressant	Immune suppressant
	Bladder carcinogen	Lipid accumulation in the liver

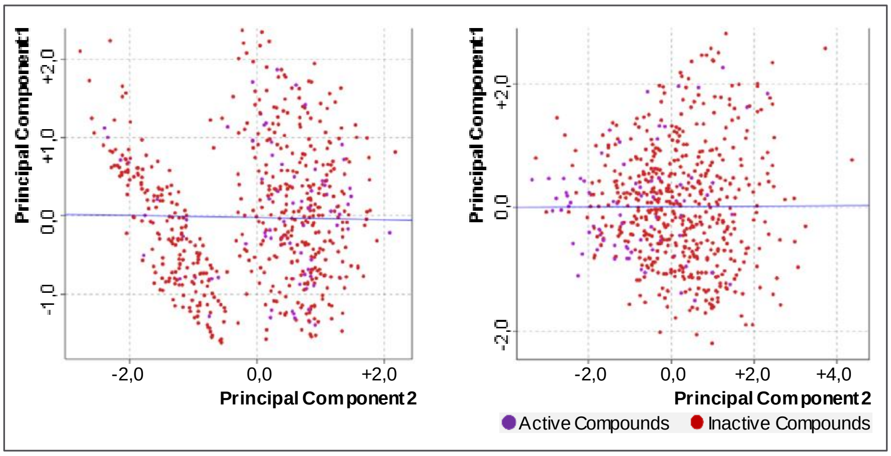


Figure 1: The molecules can be analyzed in terms of the similarity between their molecular descriptors by using the UMAP nonlinear dimensionality-reduction technique. The embedded neural representation shows an increased tendency to cluster molecules by biological activity, rather than simply on structural similarity.

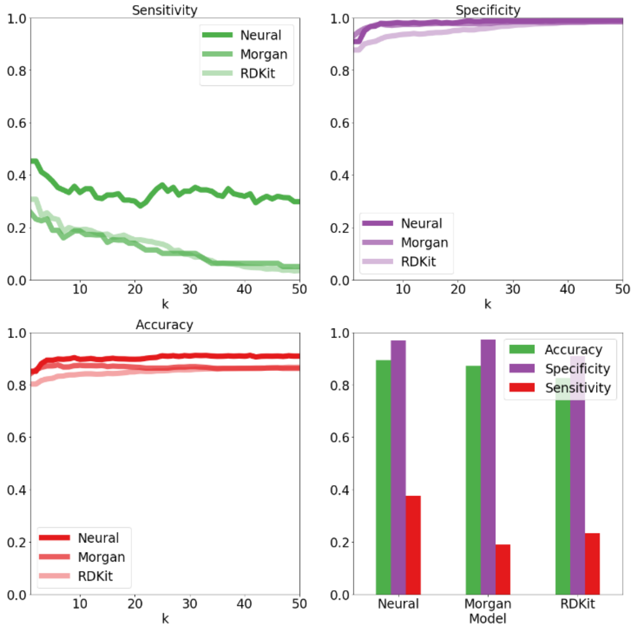


Figure 2: Performance of k-nearest neighbor classification for the three different descriptors on a separate test dataset. Visualized is the percentage accuracy, sensitivity and specificity. In general, the specificity is very low, because the dataset is imbalanced, with respect to the ratio of inactive and active compounds in the dataset. Here the results are shown exemplarily for the estrogen receptor.

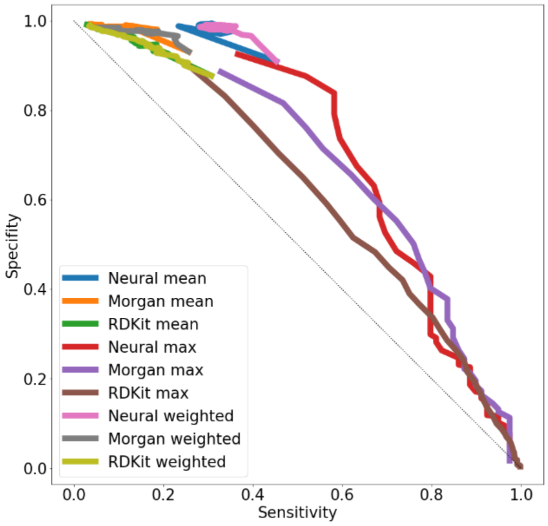
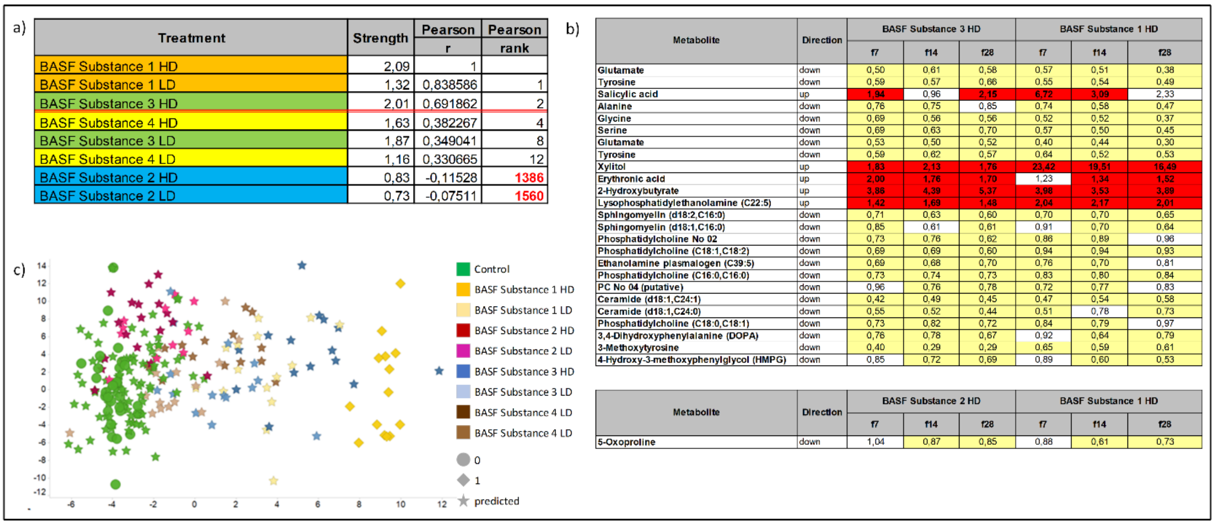


Figure 3: Evaluating the performance of combinations of distance measure and descriptor using receiver operator curves. In general, curves using the neural representation have higher areas under the curve indicating improved separation of active and inactive molecules



a)

Treatment	Strength	Pearson r	Pearson rank
BASF Substance 1 HD	2,09	1	
BASF Substance 1 LD	1,32	0,838586	1
BASF Substance 3 HD	2,01	0,691862	2
BASF Substance 4 HD	1,63	0,382267	4
BASF Substance 3 LD	1,87	0,349041	8
BASF Substance 4 LD	1,16	0,330665	12
BASF Substance 2 HD	0,83	-0,11528	1386
BASF Substance 2 LD	0,73	-0,07511	1560

b)

Metabolite	Direction	BASF Substance 3 HD			BASF Substance 1 HD		
		f7	f14	f28	f7	f14	f28
Glutamate	down	0,50	0,61	0,58	0,57	0,51	0,38
Tyrosine	down	0,59	0,57	0,66	0,55	0,54	0,49
Salicylic acid	up	1,94	0,96	2,15	6,72	3,09	2,33
Alanine	down	0,76	0,75	0,85	0,74	0,58	0,47
Glycine	down	0,69	0,56	0,56	0,52	0,52	0,37
Serine	down	0,69	0,63	0,70	0,57	0,50	0,45
Glutamate	down	0,53	0,50	0,52	0,40	0,44	0,30
Tyrosine	down	0,59	0,62	0,57	0,64	0,52	0,53
Xylitol	up	1,83	2,13	1,76	23,42	19,51	16,49
Erythronic acid	up	2,00	1,76	1,70	1,23	1,34	1,52
2-Hydroxybutyrate	up	3,86	4,39	5,37	3,98	3,53	3,89
Lysophosphatidylethanolamine (C22:5)	up	1,42	1,69	1,48	2,04	2,17	2,01
Sphingomyelin (d18:2,C16:0)	down	0,71	0,63	0,60	0,70	0,70	0,65
Sphingomyelin (d18:1,C16:0)	down	0,85	0,61	0,61	0,91	0,70	0,64
Phosphatidylcholine No 02	down	0,73	0,76	0,62	0,86	0,89	0,96
Phosphatidylcholine (C18:1,C18:2)	down	0,69	0,69	0,60	0,94	0,94	0,93
Ethanolamine plasmalogen (C39:5)	down	0,69	0,68	0,70	0,76	0,70	0,81
Phosphatidylcholine (C16:0,C16:0)	down	0,73	0,74	0,73	0,83	0,80	0,84
PC No 04 (putative)	down	0,96	0,76	0,78	0,72	0,77	0,83
Ceramide (d18:1,C24:1)	down	0,42	0,49	0,45	0,47	0,54	0,58
Ceramide (d18:1,C24:0)	down	0,55	0,52	0,44	0,51	0,78	0,73
Phosphatidylcholine (C18:0,C18:1)	down	0,73	0,82	0,72	0,84	0,79	0,97
3,4-Dihydroxyphenylalanine (DOPA)	down	0,76	0,78	0,67	0,92	0,64	0,79
3-Methoxytyrosine	down	0,40	0,29	0,29	0,65	0,59	0,61
4-Hydroxy-3-methoxyphenylglycol (HMPG)	down	0,85	0,72	0,69	0,89	0,60	0,53

Metabolite	Direction	BASF Substance 2 HD			BASF Substance 1 HD		
		f7	f14	f28	f7	f14	f28
5-Oxoproline	down	1,04	0,87	0,85	0,88	0,61	0,73

Figure 4 Comparison of metabolite profile of female rats treated with BASF Substance 1, 2, 3 and 4. a) Profile Comparison for BASF Substance 1 high dose (HD) indicating its own low dose (LD) to represent the best match directly followed by BASF Substance 3 at rank 2. BASF Substance 2 was only found at rank 1386 and 1560, respectively. The red line indicates the threshold for matching treatments (Pearson r of 0.5). b) Commonly shared metabolites of BASF Substance 1 with BASF Substance 2 and 3. A large subset of commonly changed metabolites was found for BASF Substance 1 and 3, whereas only one single metabolite was shared between BASF Substance 1 and 2. Red indicates a significant upregulation, yellow indicates a significant downregulation (p=0.05). f7,14 and 28 = data from female rats at day 7, 14 and 28. c) OPLSDA set up with the metabolic profile controls and BASF Substance 1 treated female rats. BASF Substance 3 was predicted to be the most similar to BASF Substance 1, clustering well with its LD. BASF Substance 2 is clustering more closely with the controls.