This year's workshop will focus on the concept of data uncertainty and its importance in decision-making strategies and development of predictive tools. The synthesis of overlapping data sources with varying uncertainty and reliability is critical in developing robust predictive methods, yet it is one of the aspects of model development that is often undervalued or completely disregarded. We will discuss this issue from the point of view of experimentalists, modelers and practitioners involved in safer chemical design. The three specific topic areas that will be highlighted include data uncertainty in model development, alternatives and hazard assessments, and in developing platforms for data sharing. In this session, Prof. Jakub Kostal outlines a new systematic approach for assessing data quality in skin permeation studies based on updated Klimisch scoring; Prof. James Rathman discusses a quantitative weight-of-evidence approach for estimating uncertainty and integrating alerts, read-across and QSAR; Dr. Hans Plugge focuses on data uncertainty in read-across approaches for alternatives assessments; Dr. Lauren Heine presents strategies for addressing data uncertainty in comparative hazard assessment, and Dr. Valery Tkachenko outlines challenges with data quality in chemical databases.

Jakub Kostal

Chemistry department, The George Washington University
DOT Consulting, LLC

In building predictive models for toxic endpoints, we rely on experimental data. Seemingly, having more data is considered beneficial, but is it always the case? As we continue to move into the data-driven era, modelers are tasked with finding a balance between having a large number of compounds for a training set and ensuring the studies for those compounds are of high quality. In this study, we developed a quantitative system for evaluating skin permeation data quality that allows toxicologists to build models on more reliable training sets. Transparency in reporting experimental details is key, and study parameters such as the person's age, gender, race, and skin excision location are important for ensuring *in-vitro*/*in-vivo* concordance. To guarantee reproducibility, steady-state flux in diffusion cells and homogeneity of samples must be maintained. Our system allows toxicologists to score skin permeation studies based on carefully selected parameters and select only those that fit established data-quality thresholds.

James F. Rathman

Chemical and Biomolecular Engineering, Ohio State University,
Altamira LLC
Molecular Networks GmbH

Computational safety and risk assessment relies on evaluations of diverse information sources, including structural alerts, QSAR models, and read-across for data imputation. We have been exploring the application of a decision theory approach, Dempster-Shafer theory (DST), to computational modeling of chemical toxicity. DST is an especially promising approach for addressing two key issues that have not yet been resolved in the safety and risk assessment of chemicals. The first issue is to provide quantitative and accurate estimations of the uncertainty associated with computational results. While conventional statistics allows us to account for variability in the data used to build computational models, DST allows us to also account for variability due to incomplete knowledge ("ignorance"), the fact that a mathematical model itself may not consider or properly account for all relevant effects, and is thus also a source of uncertainty. Using skin sensitization potential as an example, we illustrate how DST can be applied in QSAR modeling to provide a quantitative measure of uncertainty for each prediction generated by the model. Examples are shown for both binary classification and multi-level ordinal classification. DST is also well-suited for addressing a second key issue: the need for more rigorous techniques of combining evidence from multiple sources to arrive at a consensus decision. We describe the application of DST in a computational workflow for carcinogenicity potential that integrates results from multiple QSARs, read-across, and structural alerts. Each source is weighted in the combination-of-evidence process according to its reliability. We discuss how reliability measures can be obtained for the various types of evidence sources.

Hans Plugge

3E Company

Alternatives Assessments started out with list-based hazard assessment systems progressing through classification-based systems to raw scientific data-based systems. At each step, method uncertainty has decreased leaving eventually only the uncertainty in the raw data. Raw data quality is hard to ascertain. GLP compliance is one indicator of a quality study. Reliability indicators such as those used by ECHA add another level of comfort. Those two indicators alone will significantly limit the uncertainty in the raw data, generally eliminating outliers. Outliers that can be hard to spot are misreported data e.g. wrong units or data reported as greater than. One of the major sources of remaining uncertainty is extra/interpolation of data to fill in the missing data points. The major techniques in use today are Read Across and QSAR. Although both approaches have major uncertainties, at present, a good matrix comparison for Read Across is the best one can do for all chemicals. QSAR programs are now relatively well established for predicting physical properties of chemicals, but, when it comes to predicting toxicological properties, barring certain special cases, QSAR falls short at present, and introduces major uncertainties. Quantifying any of these uncertainties is a major headache. Hazard assessments generally use transformed data, most on a logarithmic scale. Uncertainties are thus "dampened" but are not generally delimited. Based on our 3E Green Score, we developed an approach to minimize Read-Across uncertainties using a matrix approach i.e. a minimum 2x3 matrix around the missing data. We will provide data for linear and branched aldehydes with data gaps filled based on Read-Across data. Linear aldehydes provide the best correlation, especially when compared to extrapolation from corresponding ketones. Preliminary data appears to indicate that Read-Across from branched isomers needs to be evaluated carefully. Combining all these approaches leads to a decreased uncertainty in Alternatives Assessment scores, especially for missing data points.

Valery Tkachenko

Royal Society of Chemistry

Chemical databases have been around for decades and over years there was a steady trend of rising size and complexity which still remained quite linear. In last two decades though, as a part of a digital technological revolution, we observed an unprecedented growth in sizes and types of chemical data and databases. Such qualitative change was not supported just by increasing computing power and cheapening hardware – it had its roots in appearance of principally new ecosystem in chemical data sciences – the one that rests on agile development principles, flexible licensing and open-source code. Such ecosystem though still being in its infantry is only developing standards and approaches. As consequence the questions of data models, data exchange, data formats, data quality, data control and semantic applicability are of highest importance and are far from being answered. In this presentation we will talk about the basic principles of cheminformatics and chemical databases, their reflection to the quality of the data, the brief history of ChemSpider and its value for community, the lessons learned by participation in development of the two major chemical databases (PubChem and ChemSpider) and some other products developed at the Royal Society of Chemistry and our efforts to create an open platform for chemical data.