**Data Mining and Informatics-based Approaches for New and Emerging Environmental Contaminants**

Daniel T. Chang, Michael-Rock Goldsmith, Chris M. Grulke, Peter P. Egeghy, Yu-Mei Tan,
U.S. Environmental Protection Agency, National Exposure Research Laboratory

Jade Mitchell-Blackwood
Michigan State University, College of Engineering

Corresponding author: Daniel T. Chang
Mail Code: E205-01, 109 T.W. Alexander Drive, Research Triangle Park, NC 27711
Phone: 919-541-0875
Email: chang.daniel@epa.gov

**Key Words**
Cheminformatics, ADME, Pharmacokinetics, Knowledge Discovery, QSAR, Molecular Models, Environmental Contaminants, Exposure Science

***Disclaimer:*** *The United States Environmental Protection Agency has provided administrative review and has approved for publication. The views expressed in this chapter are those of the authors and do not necessarily reflect the views or policies of the United States Environmental Protection Agency.*

New and emerging environmental contaminants are chemicals that have not been previously detected or that are being detected at levels significantly different than expected in both biological and ecological arenas (i.e., human, wildlife and environment). Many chemicals can originate from a variety of sources including consumer, agriculture, and industry as well as natural and/or anthropogenic disaster scenarios. For example, endocrine disrupting chemicals (EDCs), pharmaceuticals and personal care products (e.g., therapeutic, non-therapeutic and veterinary drugs, as well as cosmetics and fragrances) are known to be present in many of the world's water bodies and thought to originate from a variety of sources including improper disposal into municipal sewage, agribusiness and veterinary practices. The detection and quantification of these chemicals from a toxicology and exposure perspective is paramount to understanding their effects on both ecosystem and human health. EDCs act on the endocrine system and are known to alter sexual development and fertility in many vertebrate species. It is suspected that they may play a role in species population decline as well as public health issues.

Discriminating between potential contaminants and noncontaminants (e.g., EDCs vs. non-EDCs) can be an exhaustive and costly endeavor. In some cases, these methods rely on a specialized detection apparatus, discrete samples and complicated sampling techniques as well as bioethical issues in testing methods that may require the sacrifice of animals. Current high

throughput screening (HTS) efforts (i.e., toxicity *in vitro* assays) are helping to reduce some of the challenges and hurdles to testing these chemicals. However, the analysis and interpretation of results requires powerful data analytics to summarize and "make sense" of the data due to the voluminous amount generated. Informatic-based approaches like cheminformatics hold the best possibility of deciphering the barrage of data within a statistical and chemical context.

**Data mining and Informatics**

Informatics is a broad field of study encompassing computer science and information technology - from the retrieval and storage of data to the mining of patterns that exist within the stored data streams. Data mining itself is one step along a process commonly known as data or knowledge discovery (KD). Data curation and storage are critical steps in this process, but analysis and interpretation help researchers to elucidate and summarize patterns and relationships within the data itself through sophisticated algorithmic and visualization-type techniques. Some examples of these data mining techniques include the location of predetermined groups (e.g., decision tree and random forest classifiers), organization of data due to logical relationships (e.g., hierarchical, *k*-means and *k*-nearest neighbor clustering), identification of associations/dependencies (e.g., associative rule mining), and prediction of patterns based on historical data (e.g., predictive analytics). Many of these techniques are routinely applied in the retail, finance and marketing sectors (e.g., predicting consumer buying habits and trends in the market). Informatic-based approaches in the life sciences are largely dominated by chem- and bioinformatics which also employ the same techniques to understand the relationships - associations and patterns - related to chemical structure/properties and biological function, respectively. Historically, this has been done with an eye towards towards discovery of new chemicals. This discovery aspect has apparent implications in both pharmaceutical and material sciences, but the same tools and techniques are beginning to be utilized in a variety of research areas (e.g., GIS, environment- and genome-wide association studies). A visual representation using IBM's Many Eyes service (http://www-958.ibm.com) using 374 abstracts found in PubMed with the search query {"data mining"[All Fields] AND "chemical*"[All Fields] AND hasabstract[text]} highlights the relationships of concepts found in the current literature (Figure 1).

**Chemical space and cheminformatics**

Application of data mining techniques in the arena of knowledge discovery for new and emerging chemicals encompasses a wide variety of chemicals from exposure biomarkers and pesticides to drugs and EDCs. Since "chemical space" is defined by the set of all energetically stable stoichiometric combinations of atoms, nuclei and electrons, it is not difficult to imagine that the possible combinations are astronomical - easily surpassing the current list of emerging contaminants. The number of small organic chemicals alone have been estimated to be on the order of $10^{60}$. To sample and characterize this space, it would require multiple lifetimes at the current level of technology. Optimistically, this would be on the order of $10^{52}$ years assuming most HTS efforts can process ~100,000 chemicals/day! A far more efficient approach would be to apply cheminformatic-based data mining techniques to the subset of already known chemicals and arrive at a qualitative and potentially quantitative predictive framework (i.e., through clustering, classification, association and predictive analytics). This approach would

provide context for characterizing existing chemicals as well as new and emerging chemicals through a combination of molecular descriptor generation, molecular fingerprinting and predictive analytics.

The description of chemical space is largely dictated by structure-based information. For example, one could define chemical space in terms of the subset of chemical properties (i.e., molecular descriptors) that might be of interest to a particular biological activity or outcome. Molecular descriptors have a long history within cheminformatics and can be categorized as either mathematical constructs or empirical based measurements that allow one to enumerate/quantify information about a chemical - spanning the range from simple quantification of a chemical's relative partition into oil and water (log $P_{o/w}$) to more complex quantum mechanical-based descriptors that rely on the electron density of a molecule.

Molecular fingerprints, much like their name implies, are encoded structure-based information (e.g., molecular descriptors, fragments) that are ideally unique to a particular chemical. As variable or fixed-sized representations, they can encode structural keys related to both 2D and 3D molecular information. The power of molecular fingerprints is that they can be rapidly evaluated, and compared to existing fingerprints in a database thereby making similarity/dissimilarity searches trivial via standard similarity measures (i.e., Tanimoto Index). Chemical similarity is largely based on the principle that similar compounds have similar properties and hence by association chemicals can be grouped on the basis of some derived similarity in their selected molecular fingerprints (i.e., P-glycoprotein inhibitors vs. non-inhibitors). Calculated distance matrices among a database of chemicals can also aid in identifying observed structure in the data (i.e., clustering of like properties and/or biological activity).

One of the classic predictive analytic methods of cheminformatics is quantitative structure-activity relationship which seeks to find statistical correlations between a finite set of structure-based features (i.e., molecular descriptors) and their observed outcome (i.e., molecular and/or biological activity). Due to the feature selection problem (i.e., which descriptors to choose in the model) a variety of algorithms have evolved to utilize data mining techniques such as neural networks, support vector machines, and ensemble average and kernel based methods. Applicability domain issues (i.e., the relevancy and applicability of a predictive model to a wide range of chemicals) are always prevalent in such models, as they rely heavily on the available data to "train" their predictive associations. In such cases, local models that are defined by their nearest neighbors association may provide more predictive power than global models by interpolating within the data rather than extrapolating outside the data. However, these models may suffer from sparse data and/or small training sets making it difficult to accurately quantify the applicability domain.

Visualization of multiple molecules of interest within a set of prescribed descriptor dimensions can convey rapid information on chemical similarity/dissimilarity as well as general clustering of chemicals. Reduced dimensionality visualization approaches, such as three-dimensional principal component analysis (3D-PCA) plots, can provide rapid visual insights. In this case, the

similarity/dissimilarity of chemicals is based on the relative mapped positions of one molecular entity's structure-based properties with relation to another "neighboring" entity in a reduced euclidean space composed of multiple molecular descriptors. We highlight a chemographic representation based on CHEMGPS-NP (http://chemgps.bmc.uu.se) of several open-access chemical databases that illustrate the representation of multidimensional data in identifying overlaps in datasets based on similar principal components (Figure 2).

**Exposure science and pharmacokinetics**

Detection of a chemical can be suggestive to its presence in the environment. However, a chemical's presence alone does not dictate the effect on ecosystem and human health due to many determining factors. Analogously, exposure to a chemical does not necessarily mean that an untoward effect (i.e., toxicity, disease) will arise since a complex and complicated relationship exists between many determining factors including the physico-chemical properties, the concentration in the environment, the subsequent fate and transport within both biology and the environment, and the discrete exposure related behaviours (i.e., time-activity patterns) of the biological receptor (e.g., non-target wildlife species, susceptible individuals/populations). Understanding these factors is a primary concern of exposure science which seeks to understand the continuum of processes from a chemical source to a tissue dose within an organism. The range of predicted physico-chemical properties for new and emerging contaminants, however, may influence these key factors thus making efforts at determining chemical similarity and their associated properties with predictive analytics a critical step in characterizing these chemicals.

Environmental fate and transport as well as its biological analogue, pharmacokinetics/pharmacodynamics, is described by the chemical's interaction within the system. In the pharmaceutical sciences, simple pharmacokinetic-based ADME (Absorption-Distribution-Metabolism-Elimination) "rule of thumbs" are commonly used as a selective criteria in screening for drug candidates quickly and efficiently. The most famous of these is Lipinski's "Rule of Five" (RO5) and subsequent variations which seek to identify "druglikeness" in candidate compounds (i.e., orally active drugs for humans) based on its permeability/absorption into the body. Like many generalizations, it is far from perfect with many limitations due to its inability to cover all of drug space (i.e., domain of applicability issues based on four simple molecular descriptors). But as a screening tool, it was transformative in the science, successfully showing that drug permeability could be screened based on simple molecular descriptors thus narrowing down the pool of candidate drugs cheaply and efficiently.

From a human exposure perspective, ADME concepts can be used to characterize exposure potentials of chemicals based on a rate-limiting step assumption of how chemicals enter/exit the body. If we assume that ADME, a step along the source-to-outcome continuum, describes the biological process whereby a chemical trespasses the body's barrier (absorption), is metabolized, distributed and exits the body (elimination), then a simplified binary (fast/slow) diagram can illustrate the effect on exposure-dose relationships to categorize 16 unique scenarios - $2^4$ possible combinations - or dose categorizations (Figure 3). In these scenarios, two dominant exposure-dose themes are observed: 1) absorption limited (AL) via slow

absorption and 2) elimination limited (EL) via fast absorption into the body. In this thought experiment, one could flag potential chemicals of concern based on their ability to enter quickly and exit slowly. For example, dose categories 13, 14, 15 and 16 would have the highest concerns given that elimination is slow and absorption is fast. Conversely, dose categories 1, 2, 3 and 4 would have the lowest concerns based on slow absorption and fast elimination. Assuming simple metabolic clearance (i.e., no metabolic activation of toxicity pathways), inclusion of metabolism would delineate each category further by a faster/slower metabolism which would result in quicker/slower clearance of a chemical thus reducing/increasing its dose at a target tissue. Since all steps in the ADME process can be influenced by its physico-chemical properties, generic pharmacokinetic modeling should be used when possible to give context to the relative mappings of potential ADME behaviours alongside their predicted molecular descriptors.

**Conclusions**

Cheminformatics techniques are typically much less intensive to apply, but provide key insights into the nature of chemicals - especially in the context of knowledge discovery. For many contaminants, there is a paucity of data availability to parameterize models and perform the necessary risk assessment studies. Data mining and informatics-based approaches allow us to induce predictive models as well as intuit potential chemical similarities/dissimilarities of new and emerging contaminants to the environment and to public health. However, care should also be taken when considering the exposure-dose relationships, especially with respect to pharmacokinetic-based ADME concepts. As more information via HTS studies becomes available, the associative power of these predictive models should become richer and more detailed improving upon the current state of the science. Ultimately, the ability to rapidly characterize the presence of new and emerging chemicals as well as their effects on individuals, populations, and ecosystems will have beneficial implications for both exposure risk assessment and risk mitigation.

**Further Reading**

BACKGROUND: CHEMISTRY, CHEMINFORMATICS, QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIPS, ENVIRONMENTAL CONTAMINANTS, EXPOSURE SCIENCE, BIOMARKERS

**For more information see these additional references:**

J Larsson *et al.*, ChemGPS-NP: Tuned for Navigation in Biologically Relevant Chemical Space, *J Nat Prod.*, 70(5):789-794, 2007.

J Rosén *et al.*, ChemGPS-NPWeb: chemical space navigation online, *J Comput Aided Mol Des.*, 23(4):253-259, 2009.

TI Oprea and J. Gottfries, Chemography: the art of navigating in chemical space. *J Comb Chem* 3(2):157-66, 2001.

CA Lipinski, *et al.*, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* 23: 3-25, 1997.

U Fayyad *et al.*, "From Data Mining to Knowledge Discovery in Databases". AI Magazine, 17(3):37-54, 1996.

Toxicity Testing in the 21st Century: A Vision and a Strategy: The National Academies Press, 2007.

Exposure Science in the 21st Century: A Vision and a Strategy: The National Academies Press., 2012.

http://www.epa.gov/ppcp

**Bibliography:**

MF Rahman, *et al.*, Endocrine disrupting compounds (EDCs) and pharmaceuticals and personal care products (PPCPs) in the aquatic environment: implications for the drinking water industry and global environmental health. *J Water Health* 7(2):224-43, 2009.

CG Daughton, Pharmaceuticals in the environment: sources and their management. In *Analysis, Fate and Removal of Pharmaceuticals in the Water Cycle*. M. Petrovic and D. Barcelo, Elsevier Science. Volume 50: 1-58, 2007.

Dobson CM, Chemical Space and Biology, *Nature*, 432:824–828, 2004, doi:10.1038/nature03192

DT Chang *et al.*, *In Silico* Strategies for Modeling Stereoselective Metabolism of Pyrethroids, In *Parameters for Pesticide QSAR and PBPK/PD models for Human Risk Assessment*, JB Knaak, C Timchalk and R Tornero-Velez, American Chemical Society, 1099: 245-269, 2012.

MR Goldsmith *et al.*, Informing Mechanistic Toxicology with Computational Molecular Models, In *Computational Toxicology: Methods in Molecular Biology*, B. Reisfeld and AN Mayeno, Human Press, 929:139-165, 2012.

Y Tan *et al.*, Reconstructing human exposures using biomarkers and other “clues”. *J Toxicol Environ Health B Crit Rev*, 15(1):22-38, 2012.

Egeghy, PP, et al., The exposure data landscape for manufactured chemicals. *Sci Total Environ* 414:159-66, 2012.

Figure Captions:

Figure 1. IBM Many Eyes Phrase Net visualization of 374 abstracts queried from PubMed with {"data mining"[All Fields] AND "chemical*"[All Fields] AND hasabstract[text]} illustrating the relationship of key concepts within data mining.

Figure 2. CHEMGPS-NP representation showing the overlap of chemicals within 3 publicly available databases and literature PBPK chemical-specific PBPK models through 2010. PBPK model chemicals queried from literature up to 2010; NHANES IV (http://www.cdc.gov/nchs/nhanes.htm) chemicals; USDA-PDP (http://www.ams.usda.gov/AMSv1.0/pdp) chemicals; ToxCast™ (http://www.epa.gov/ncct/toxcast/) Phase 1 chemicals. The first 4 dimensions of the principal component analysis are plotted.

Figure 3. Sixteen dose categories based on a hypothetical binary (fast/slow) stepwise pharmacokinetic scenario considering ADME only along the source-to-outcome continuum.

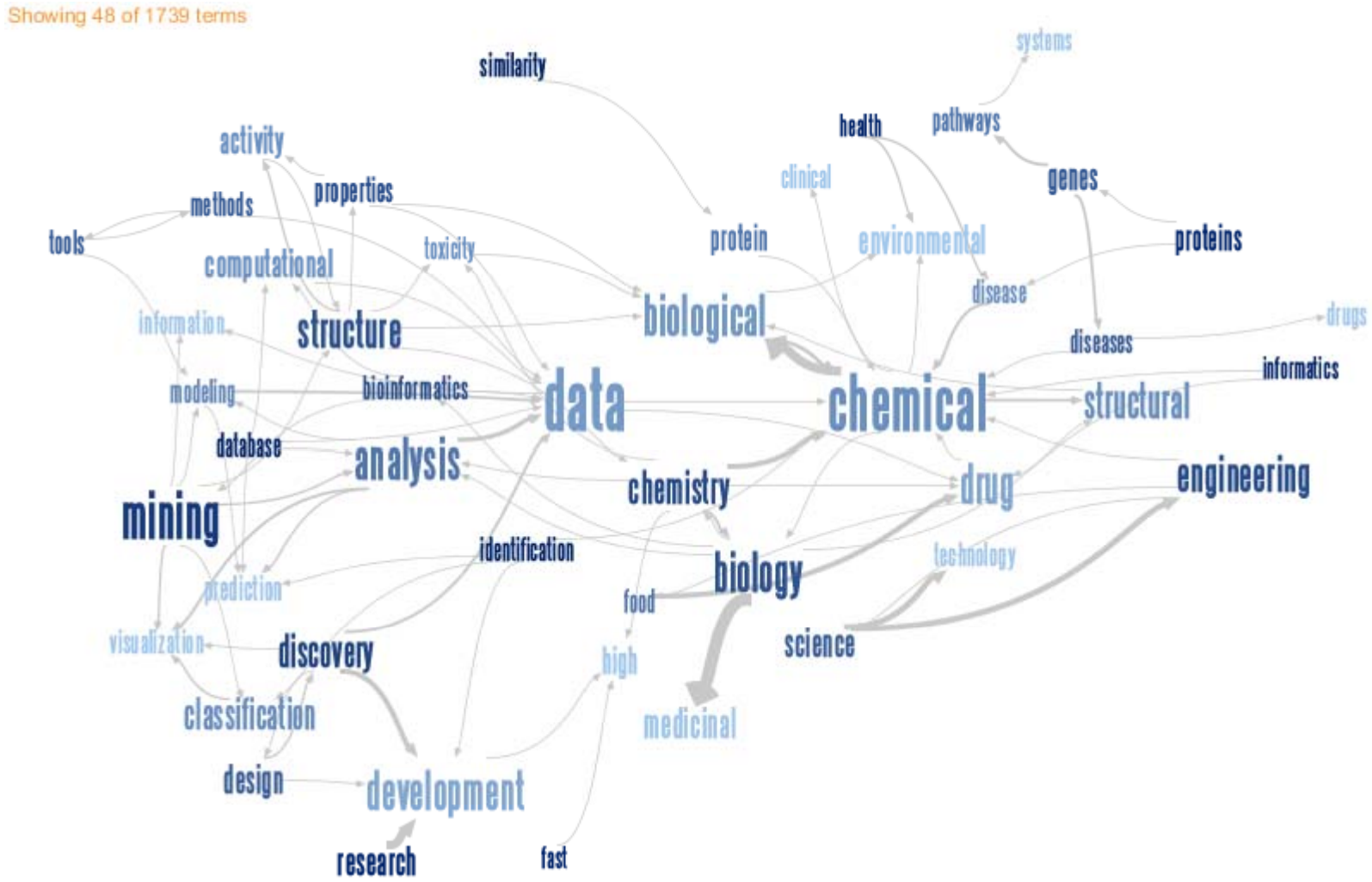
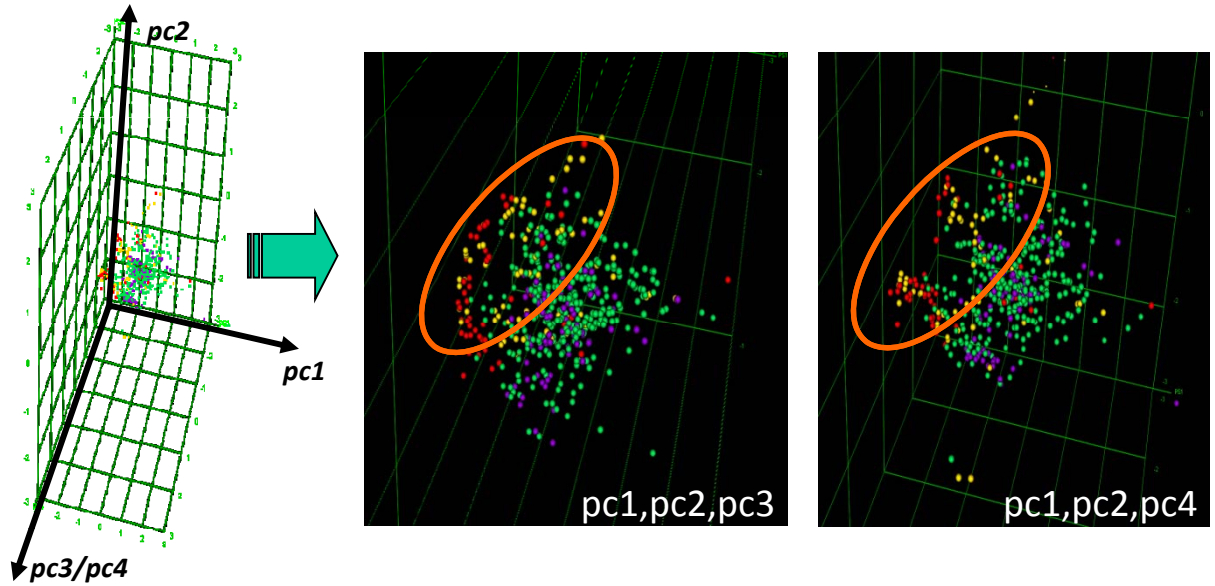

Showing 48 of 1739 terms
similarity
systems
health
pathways
activity
clinical
genes
properties
methods
tools
protein
environmental
proteins
computational
toxicity
disease
biological
information
structure
drugs
diseases
informatics
modeling
bioinformatics
data
chemical
structural
database
analysis
chemistry
drug
engineering
mining
identification
technology
biology
prediction
food
science
visualization
discovery
high
classification
medicinal
design
development
research
fast

Figure 1.




PBPK space
NHANES biomarkers
USDA-PDP
ToxCast 320

***– Principal component 1 (pc1): size, shape and polarizability***

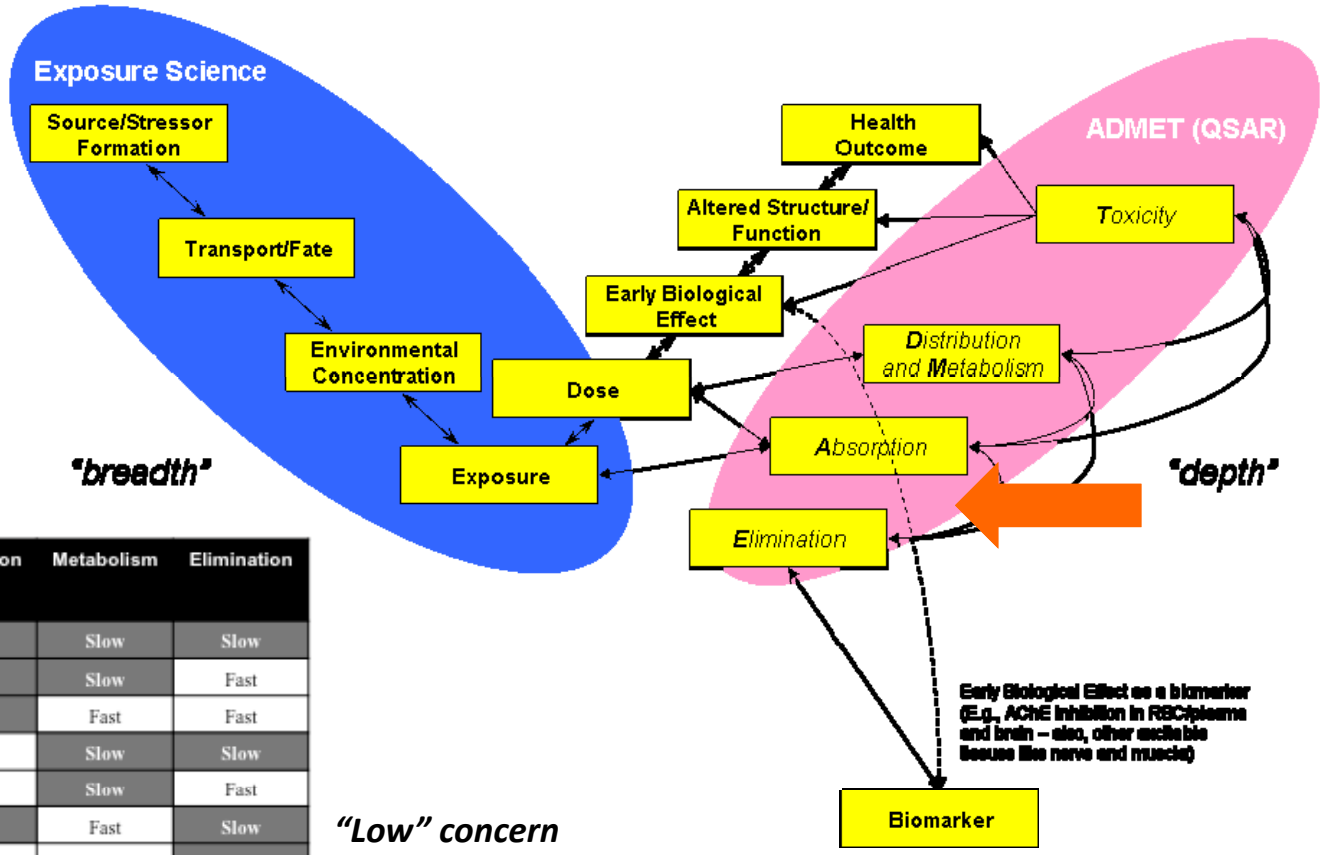
***– Principal component 2 (pc2): aromatic and conjugation related properties***

***– Principal component 3 (pc3): lipophilicity, polarity, and H-bond capacity***

***– Principal component 4 (pc4): molecular flexibility and rigidity***

Figure 2.

**Exposure makes the dose…**




| Scenario | Dose Cat | | Absorption (route-specific) | Distribution | Metabolism | Elimination |
|---|---|---|---|---|---|---|
| 1 | 8 | AL | Slow | Slow | Slow | Slow |
| 2 | 4 | AL | Slow | Slow | Slow | Fast |
| 3 | 2 | AL | Slow | Slow | Fast | Fast |
| 4 | 7 | AL | Slow | Fast | Slow | Slow |
| 5 | 3 | AL | Slow | Fast | Slow | Fast |
| 6 | 6 | AL | Slow | Slow | Fast | Slow |
| 7 | 5 | AL | Slow | Fast | Fast | Slow |
| 8 | 1 | AL | Slow | Fast | Fast | Fast |
| 9 | 16 | EL | Fast | Slow | Slow | Slow |
| 10 | 12 | EL | Fast | Slow | Slow | Fast |
| 11 | 11 | EL | Fast | Fast | Slow | Fast |
| 12 | 14 | EL | Fast | Slow | Fast | Slow |
| 13 | 10 | EL | Fast | Slow | Fast | Fast |
| 14 | 15 | EL | Fast | Fast | Slow | Slow |
| 15 | 13 | EL | Fast | Fast | Fast | Slow |
| 16 | 9 | EL | Fast | Fast | Fast | Fast |

*"Low" concern*

*vs.*

*"High" concern*

**…ADME determines whether or not the dose is a poison. (*or whether or not there is a dose!*)**

Figure 3.