Modeling Chronic Toxicity: A comparison of experimental variability with read across predictions

Christoph Helma[1], David Vorgrimmler[1], Denis Gebele[1], Martin Gütlein[2], Benoit Schilter[3], Elena Lo Piparo[3]

E-mail:

[1] in silico toxicology gmbh, Basel, Switzerland

[2] Inst. f. Computer Science, Johannes Gutenberg Universität Mainz, Germany

[3] Chemical Food Safety Group, Nestlé Research Center, Lausanne, Switzerland

Introduction

Christoph + Elena + Benoit

The main objectives of this study are

- to investigate the experimental variability of LOAEL data
- develop predictive model for lowest observed effect levels
- compare the performance of model predictions with experimental variability

Materials and Methods

Datasets

Mazzatorta dataset

Just referred to the paper 2008.

Swiss Federal Office dataset

Elena + Swiss Federal Office contribution (input)

Only rat LOAEL values were used for the current investigation, because they correspond directly to the Mazzatorta dataset.

Preprocessing

Christoph

Chemical structures in both datasets are represented as SMILES strings (Weininger 1988). Syntactically incorrect and missing SMILES were generated from other identifiers (e.g names, CAS numbers) when possible. Studies with undefined ("0") or empty LOAEL entries were removed for this study.

Algorithms

Christoph

For this study we are using the modular lazar (*la*zy *s*tructure *a*ctivity *r*elationships) framework (Maunz et al. 2013) for model development and validation.

lazar follows the following basic workflow: For a given chemical structure it searches in a database for similar structures (neighbors) with experimental data, builds a local (Q)SAR model with these neighbors and uses this model to predict the unknown activity of the query compound. This procedure resembles an automated version of *read across* predictions in toxicology, in machine learning terms it would be classified as a *k-nearest-neighbor* algorithm.

Apart from this basic workflow lazar is completely modular and allows the researcher to use any algorithm for neighbor identification and local (Q)SAR modelling. Within this study we are using the following algorithms:

Neighbor identification

Christoph

Similarity calculations are based on MolPrint2D fingerprints (Bender et al. 2004) from the OpenBabel chemoinformatics library (OBoyle et al. 2011).

The MolPrint2D fingerprint uses atom environments as molecular representation, which resemble basically the chemical concept of functional groups. For each atom in a molecule it represents the chemical environment with the atom types of connected atoms.

The main advantage of MolPrint2D fingerprints over fingerprints with predefined substructures (such as OpenBabel FP3, FP4 or MACCs fingerprints) is that it may capture substructures of toxicological relevance that are not included in predefined substructure lists. Preliminary experiments have shown that predictions with MolPrint2D fingerprints are indeed more accurate than fingerprints with predefined substructures.

From MolPrint2D fingerprints we can construct a feature vector with all atom environments of a compound, which can be used to calculate chemical similarities.

The chemical similarity between two compounds is expressed as the proportion between atom environments common in both structures and the total number of atom environments

(Jaccard/Tanimoto index (1)).

(1) $sim = \frac{|A \cap B|}{|A \cup B|}$, A atom environments of compound A, B atom environments of compound B.

Local (Q)SAR models

Christoph

As soon as neighbors for a query compound have been identified, we can use their experimental LOAEL values to predict the activity of the untested compound. In this case we are using the weighted mean of the neighbors LOAEL values, where the contribution of each neighbor is weighted by its similarity to the query compound.

Validation

Christoph

Results

Dataset comparison

Christoph + Elena

The main objective of this section is to compare the content of both databases in terms of structural composition and LOAEL values, to estimate the experimental variability of LOAEL values and to establish a baseline for evaluating prediction performance.

Applicability domain

Ches-Mapper analysis

Martin

CheS-Mapper (Chemical Space Mapping and Visualization in 3D, http://ches-mapper.org/, (Gutlein, Karwath, and Kramer 2012)) can be used to analyze the relationship between the structure of chemical compounds, their physico-chemical properties, and biological or toxic effects. CheS-Mapper embeds a dataset into 3D space, such that compounds with similar feature values are close to each other. The following two screenshots visualise the comparison. The datasets are embeded into 3D Space based on structural fragments from three Smart list (OpenBabel FP3, OpenBabel FP4 and OpenBabel MACCS).

Distribution of functional groups

Christoph

Figure 1 shows the frequency of selected functional groups in both datasets. A complete table for 138 functional groups from OpenBabel FP4 fingerprints can be found in the appendix.

Experimental variability versus prediction uncertainty

Christoph

Duplicated LOAEL values can be found in both datasets and there is a substantial overlap of compounds, with LOAEL values in both datasets.

Intra dataset variability

The Mazzatorta dataset has 562 LOAEL values with 439 unique structures, the Swiss Federal Office dataset has 493 rat LOAEL values with 381 unique structures. Figure 2 shows the intra-dataset variability, where each vertical line represents a single compound and each

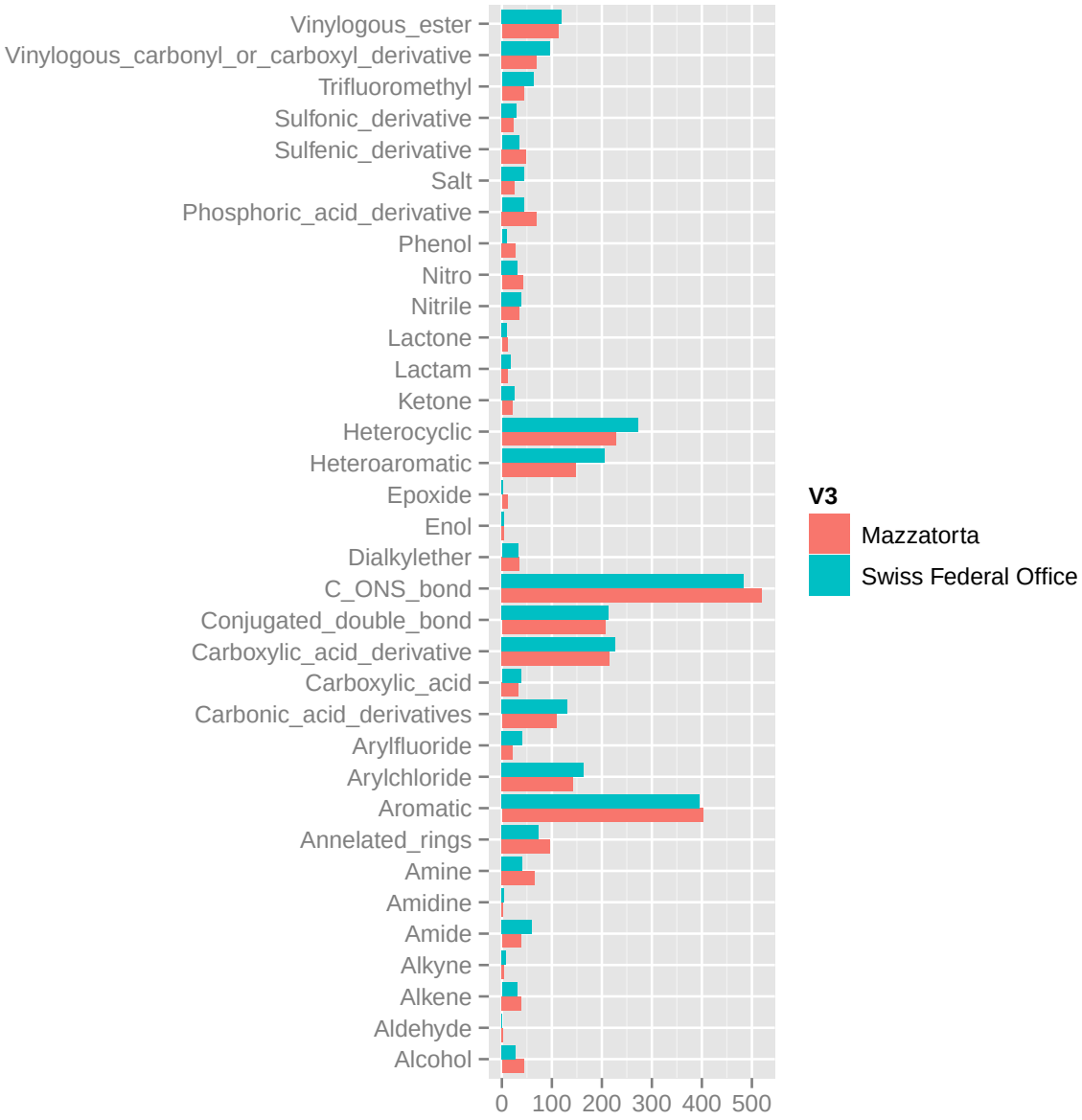

Figure 1: Frequency of functional groups.

dot represents an individual LOAEL value. The experimental variance of LOAEL values is similar in both datasets (p-value: 0.48).

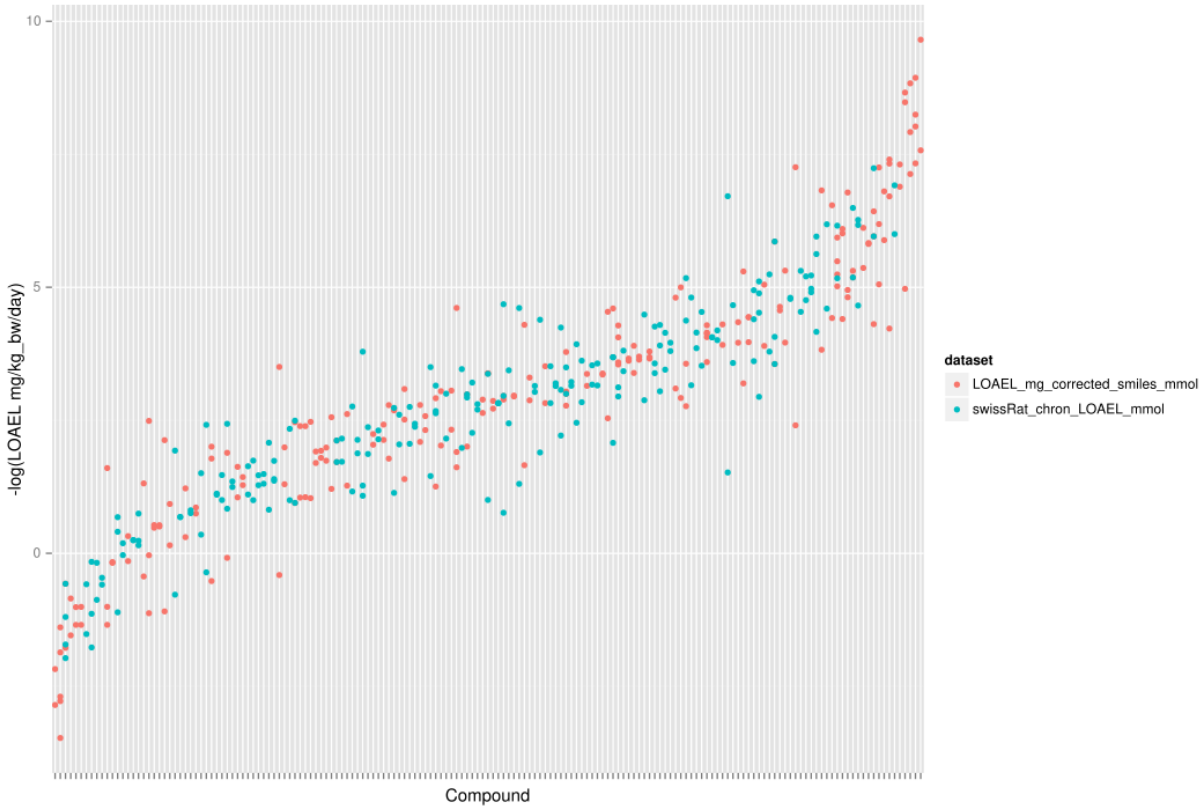

Figure 2: Intra dataset variability: Each vertical line represents a compound, dots are individual LOAEL values.

Inter dataset variability

Figure 3 shows the same situation for the combination of the Mazzatorta and Swiss Federal Office datasets. Obviously the experimental variability is larger than for individual datasets.

LOAEL correlation between datasets

Figure 4 depicts the correlation between LOAEL data from both datasets (using means for multiple measurements). Correlation analysis shows a significant correlation with r^2: 0.61, RMSE: 1.22, MAE: 0.80

Figure 3: Inter dataset variability

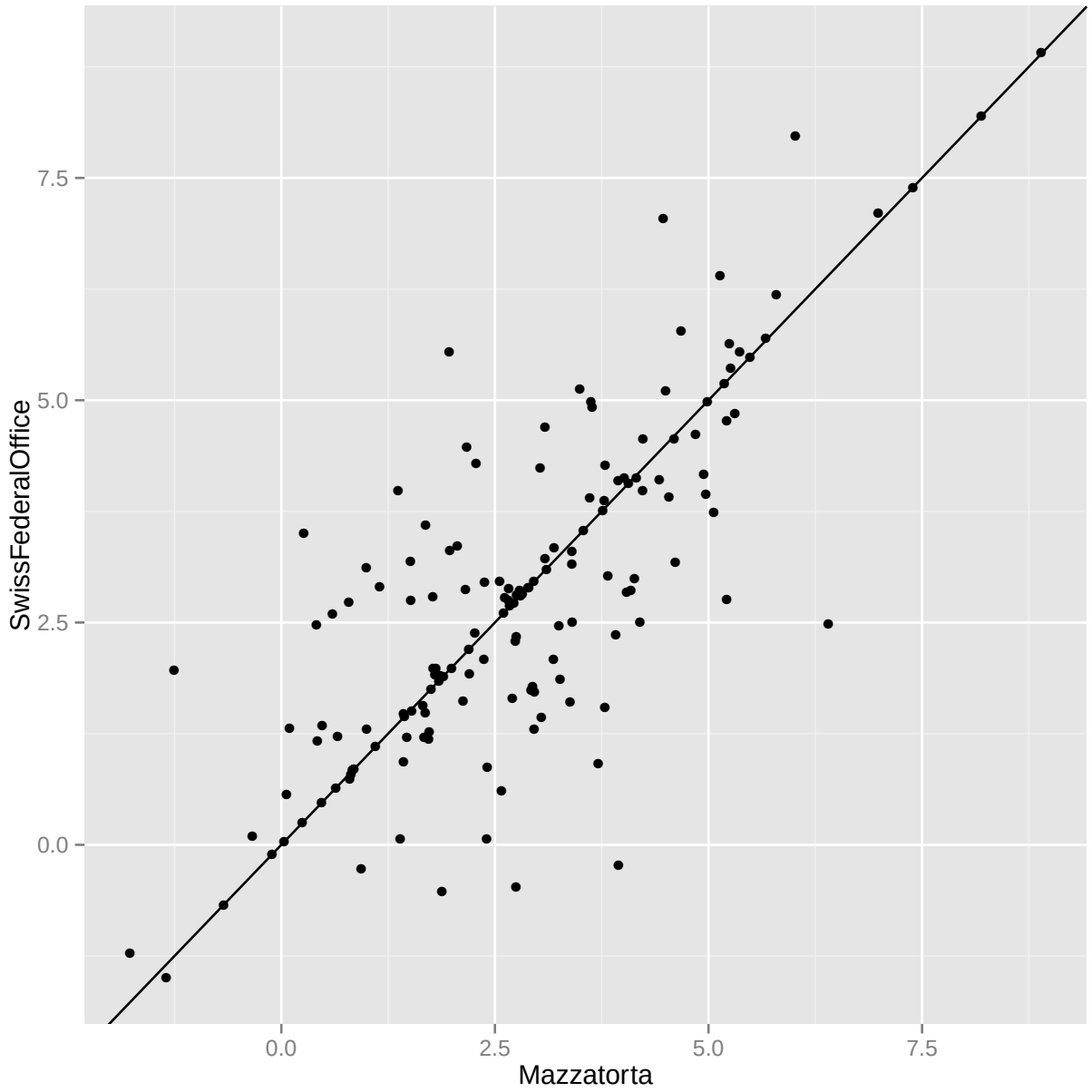

Figure 4: LOAEL correlation

Local (Q)SAR models

Christoph

Discussion

Elena + Benoit

Summary

References

Bender, Andreas, Hamse Y. Mussa, and Robert C. Glen, and Stephan Reiling. 2004. "Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier." *Journal of Chemical Information and Computer Sciences* 44 (1): 170–78. doi:10.1021/ci034207y.

Gütlein, Martin, Andreas Karwath, and Stefan Kramer. 2012. "CheS-Mapper - Chemical Space Mapping and Visualization in 3D." *Journal of Cheminformatics* 4 (1): 7. doi:10.1186/1758-2946-4-7.

Maunz, Andreas, Martin Gütlein, Micha Rautenberg, David Vorgrimmler, Denis Gebele, and Christoph Helma. 2013. "Lazar: A Modular Predictive Toxicology Framework." *Frontiers in Pharmacology* 4. Frontiers Media SA. doi:10.3389/fphar.2013.00038.

OBoyle, Noel M, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and Geoffrey R Hutchison. 2011. "Open Babel: An Open Chemical Toolbox." *Journal of*

Cheminformatics 3 (1). Springer Science; Business Media: 33. doi:10.1186/1758-2946-3-33.

Weininger, David. 1988. "SMILES, a Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules." *Journal of Chemical Information and Computer Sciences* 28 (1): 31–36. doi:10.1021/ci00057a005.