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 quality and reproducibility of (Q)SAR and read-across predictions is a controversial topic in the toxicological risk-assessment community. Although model predictions can be validated with various procedures it is rarely possible to put the results into the context of experimental variability, because replicate experiments are rarely available.

With missing information about the variability of experimental toxicity data it is hard to judge the performance of predictive models and it is tempting for model developments to use aggressive model optimisation methods that lead to impressive validation results, but also to

overfitted models with little practical relevance.

In this study we intent to compare model predictions with experimental variability with chronic oral rat lowest adverse effect levels (LOAEL) as toxicity endpoint. We are using two datasets, one from (Mazzatorta et al. 2008) (*Mazzatorta* dataset) and one from the Swiss Federal Office of TODO (*Swiss Federal Office* dataset).

Elena: do you have a reference and the name of the department?

155 compounds are common in both datasets and we use them as a test set in our investigation. For this test set we will

- compare the structural diversity of both datasets
- compare the LOAEL values in both datasets
- build prediction models based on the Mazzatorta, Swiss Federal Office datasets and a combination of both
- predict LOAELs of the training set
- compare predictions with experimental variability

With this investigation we also want to support the idea of reproducible research, by providing all datasets and programs that have been used to generate this manuscript under a TODO license.

A self-contained docker image with all program dependencies required for the reproduction of these results is available from TODO.

Source code and datasets for the reproduction of this manuscript can be downloaded from the GitHub repository TODO. The lazar framework (Maunz et al. 2013) is also available under a GPL License from https://github.com/opentox/lazar.

TODO: github tags

Elena: please check if this is publication strategy is ok for the Swiss Federal Office

Materials and Methods

Datasets

Mazzatorta dataset

The first dataset (*Mazzatorta* dataset for further reference) originates from the publication of (Mazzatorta et al. 2008). It contains chronic (> 180 days) lowest observed effect levels (LOAEL) for rats (*Rattus norvegicus*) after oral (gavage, diet, drinking water) administration. The Mazzatorta dataset consists of 567 LOAEL values for 445 unique chemical structures.

Swiss Federal Office dataset

Elena + Swiss Federal Office contribution (input)

The Swiss Federal Office dataset consists of 493 LOAEL values for 381 unique chemical structures.

Preprocessing

Chemical structures in both datasets were initially represented as SMILES strings (Weininger 1988). Syntactically incorrect and missing SMILES were generated from other identifiers (e.g names, CAS numbers). Unique smiles from the OpenBabel library (OBoyle et al. 2011) were used for the identification of duplicated structures.

Studies with undefined or empty LOAEL entries were removed from the datasets. LOAEL values were converted to mmol/kg_bw/day units. For prediction, validation and visualisation purposes -log10 transformations are used.

David: please check if we have missed something

Derived datasets

Two derived datasets were obtained from the original datasets:

The *test* dataset contains data of compounds that occur in both datasets. Exact duplications of LOAEL values were removed, because it is very likely that they originate from the same study. The test dataset has 391 LOAEL values for 155 unique chemical structures.

The *combined* dataset is the union of the Mazzatorta and the Swiss Federal Office dataset and it is used to build predictive models. Exact LOAEL duplications were removed, as for the test dataset. The combined dataset has 1014 LOAEL values for 671 unique chemical structures.

Algorithms

In 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 lazar

- searches in a database for similar structures (*neighbors*) with experimental data,
- builds a local QSAR 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 similarity searches and local QSAR modelling. Within this study we are using the following algorithms:

Neighbor identification

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 using the atom types of connected atoms.

MolPrint2D fingerprints are generated dynamically from chemical structures and do not rely on predefined lists of fragments (such as OpenBabel FP3, FP4 or MACCs fingerprints or lists of toxocophores/toxicophobes). This has the advantage the they may capture substructures of toxicological relevance that are not included in other fingerprints. Preliminary experiments have shown that predictions with MolPrint2D fingerprints are indeed more accurate than other OpenBabel fingerprints.

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 A and B is expressed as the proportion between atom environments common in both structures $A \cap B$ and the total number of atom environments $A \cup B$ (Jaccard/Tanimoto index, eq. 1).

$$sim = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

A threshold of $sim < 0.1$ is used for the identification of neighbors for local QSAR models. Compounds with the same structure as the query structure are eliminated from the neighbors to obtain an unbiased prediction.

Local QSAR models and predictions

Only similar compounds (*neighbors*) are used for local QSAR models. In this investigation we are using a weighted partial least squares regression (PLS) algorithm for the prediction of quantitative properties. First all fingerprint features with identical values across all neighbors are removed. The reamining set of features is used as descriptors for creating a local weighted PLS model with atom environments as descriptors and model similarities as weights. The `plsr` function of the `pls` R package (Mevik, Wehrens, and Liland 2015) is used for this purpose. Finally the local PLS model is applied to predict the activity of the query compound.

If PLS modelling or prediction fails, the program resorts to 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

Two types of validations are used within this study:

For the comparison of experimental variability with predictive accuracies we are using a test set of compounds that occur in both datasets. The *Mazzatorta*, *Swiss Federal Office* and *combined* datasets are used as training data for read across predictions. In order to obtain unbiased predictions *all* information from the test compound is removed from the training set prior to predictions. This is hardcoded into the prediction algorithm in order to prevent validation errors.

TODO: treatment of duplicates

In addition traditional 10-fold crossvalidation results are provided.

Christoph: check if these specifications have changed at submission

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/, Gütlein, 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

fig. 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.

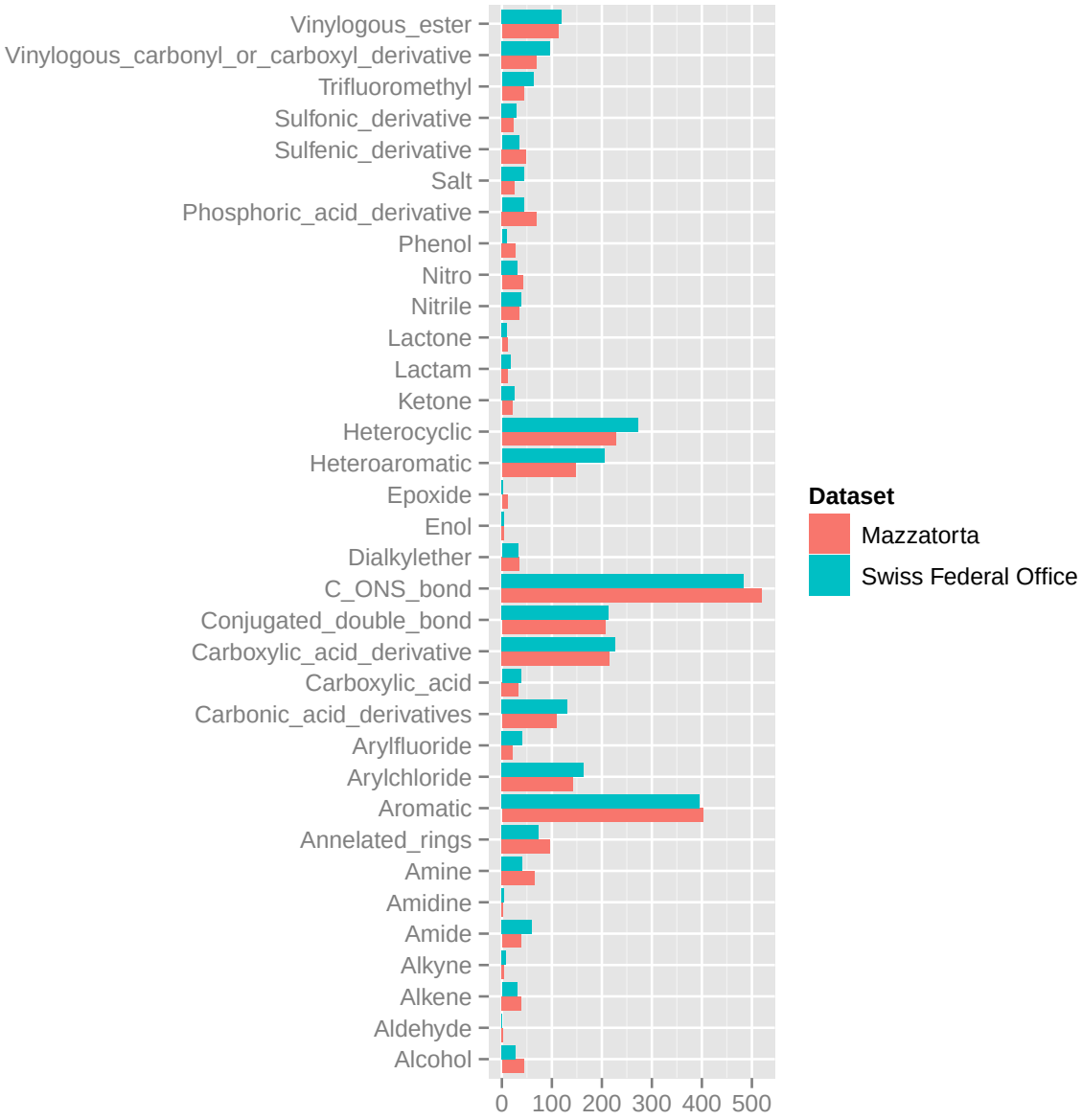

Figure 1: Frequency of functional groups.

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

TODO: read data from files

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. fig. **??** shows the intra-dataset variability, where each vertical line represents a single compound and each dot represents an individual LOAEL value. The experimental variance of LOAEL values is similar in both datasets (p-value: 0.48).

Inter dataset variability

fig. **??** 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

fig. **??** depicts the correlation between LOAEL data from both datasets (using means for multiple measurements). Identical values were removed from analysis.

Correlation analysis shows a significant correlation (p-value $<$ 2.2e-16) with r^2: 0.58, RMSE: 1.3

Local QSAR models

Christoph

In order to compare the perfomance of in silico models with experimental variability we are using compounds that occur in both datasets as a test set (155 compounds, -1 measurements).

The Mazzatorta, the Swiss Federal Office dataset and a combined dataset were used as training data. Predictions for the test set compounds were made after eliminating all information from the test compound from the corresponding training dataset. tbl. 1 summarizes the results:

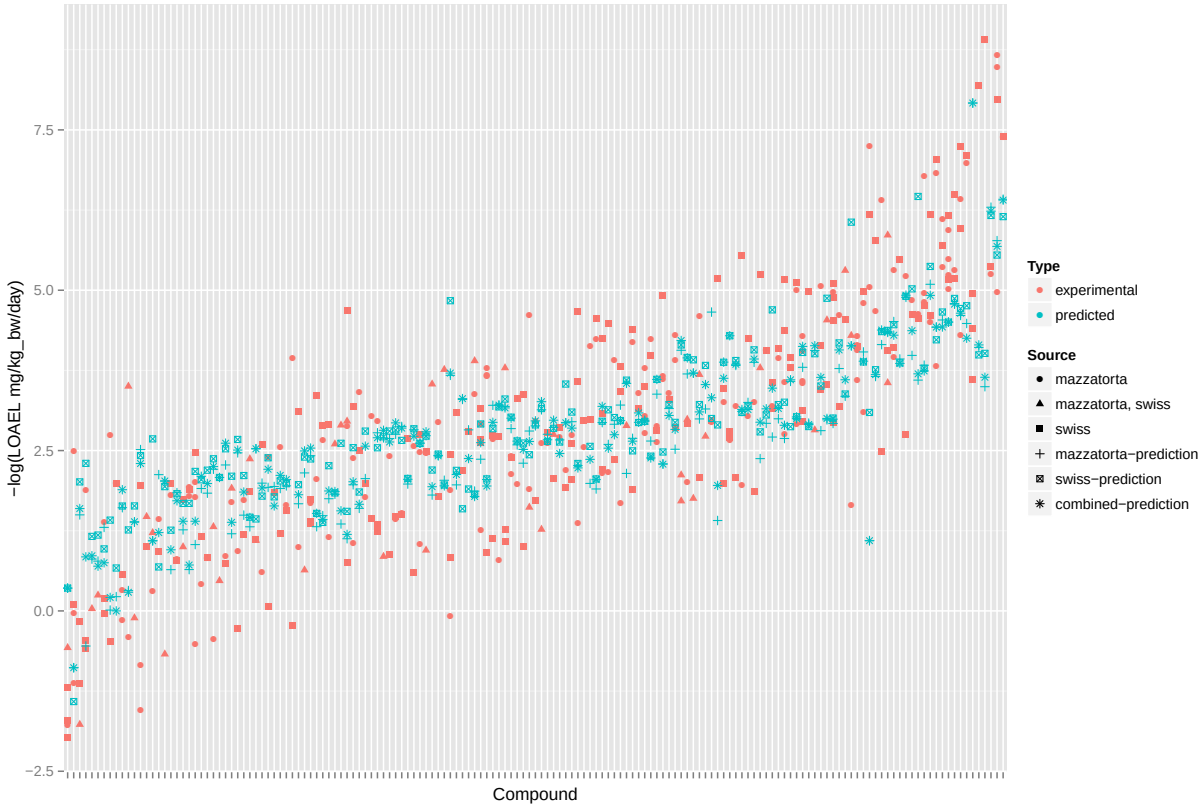

Figure 2: Comparison of experimental with predicted LOAEL values, each vertical line represents a compound.

Table 1: Comparison of model predictions with experimental variability.

Training data	r^2	RMSE
Experimental	0.58	1.3
Mazzatorta	0.38	1.49
Swiss Federal Office	0.38	1.47
Combined	0.38	1.47

Traditional 10-fold cross-validation results are summarised in tbl. 2:

Table 2: 10-fold crossvalidation results

Training dataset	r^2	RMSE
Mazzatorta	0.38	2.01
Swiss Federal Office	0.3	1.67
Combined	0.38	1.81

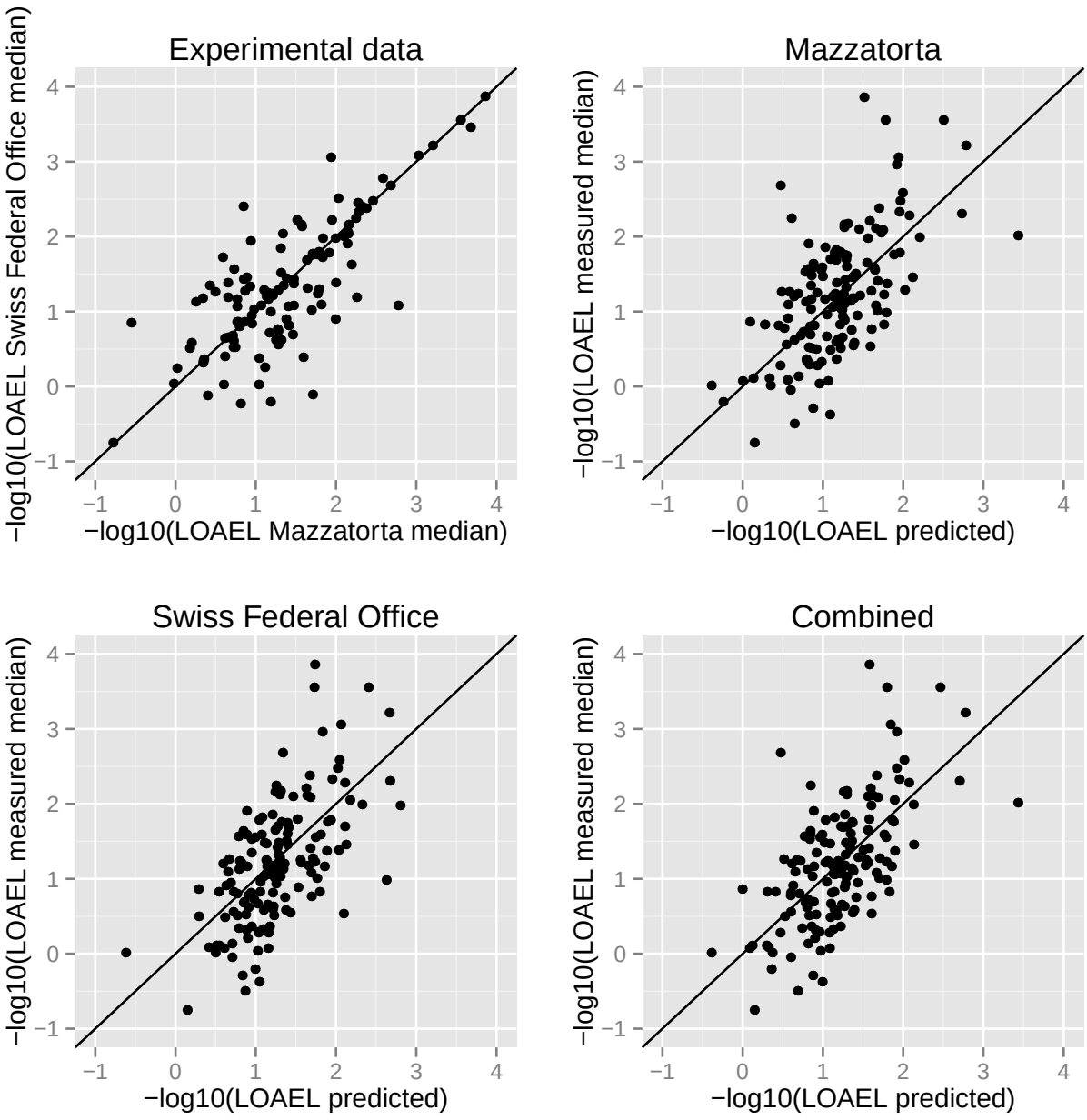

Figure 3: Correlation of experimental with predicted LOAEL values (test set)

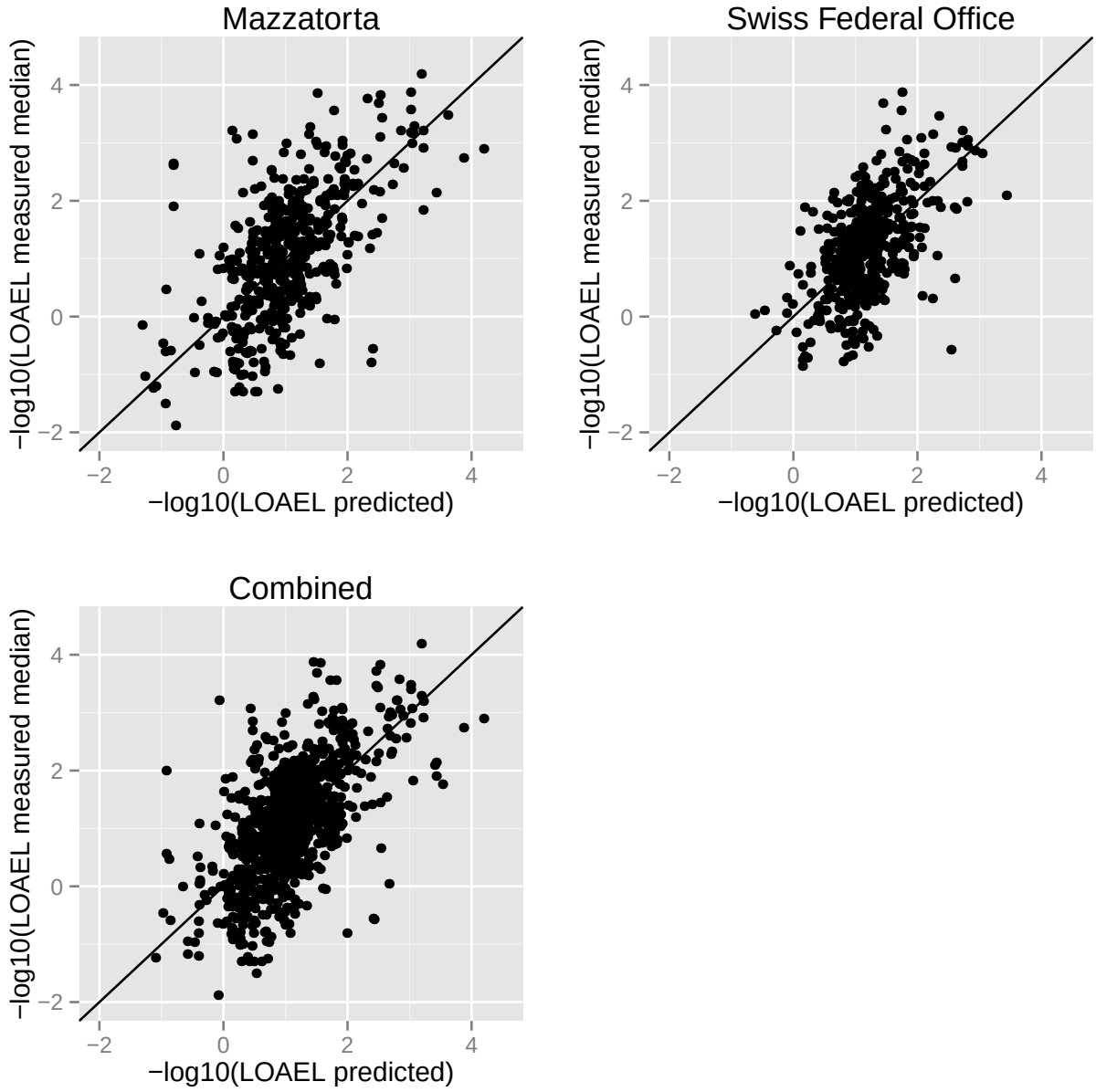

Figure 4: Correlation of experimental with predicted LOAEL values (10-fold crossvalidation)

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.

Mazzatorta, Paolo, Manuel Dominguez Estevez, Myriam Coulet, and Benoit Schilter. 2008. "Modeling Oral Rat Chronic Toxicity." *Journal of Chemical Information and Modeling* 48 (10): 1949–54. doi:10.1021/ci8001974.

Mevik, Bjørn-Helge, Ron Wehrens, and Kristian Hovde Liland. 2015. *Pls: Partial Least Squares and Principal Component Regression*. https://CRAN.R-project.org/package=pls.

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 and 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.