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

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 375 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 998 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, Equation 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 `pls` method from the `caret` R package (Kuhn 2008) 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.

default settings for tuning

Applicability domain

Christoph: TODO

Prediction intervals were obtained from the `predict` function.

Validation

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 procedure is hardcoded into the prediction algorithm

in order to prevent validation errors. Traditional 10-fold crossvalidation results are provided as additional information for all three models.

TODO: treatment of duplicates

Christoph: check if these specifications have changed at submission

Results

Dataset comparison

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.

Ches-Mapper analysis

We applied the visualization tool CheS-Mapper (Chemical Space Mapping and Visualization in 3D, http://ches-mapper.org, Gütlein, Karwath, and Kramer (2012)) to compare both datasets. CheS-Mapper can be used to analyze the relationship between the structure of chemical compounds, their physico-chemical properties, and biological or toxic effects. It embeds a dataset into 3D space, such that compounds with similar feature values are close to each other. CheS-Mapper is generic and can be employed with different kinds of features. Figure 1 shows an embedding that is based on physico-chemical (PC) descriptors: we determined that both datasets have very similar PC feature values.

We extended CheS-Mapper with a functionality to mine the same MolPrint2D features that are utilized for model building in this work. Applying a minimum frequency of 3 yields 760 distinguished MolPrint2D fragments for the composed dataset of 671 unique compounds.

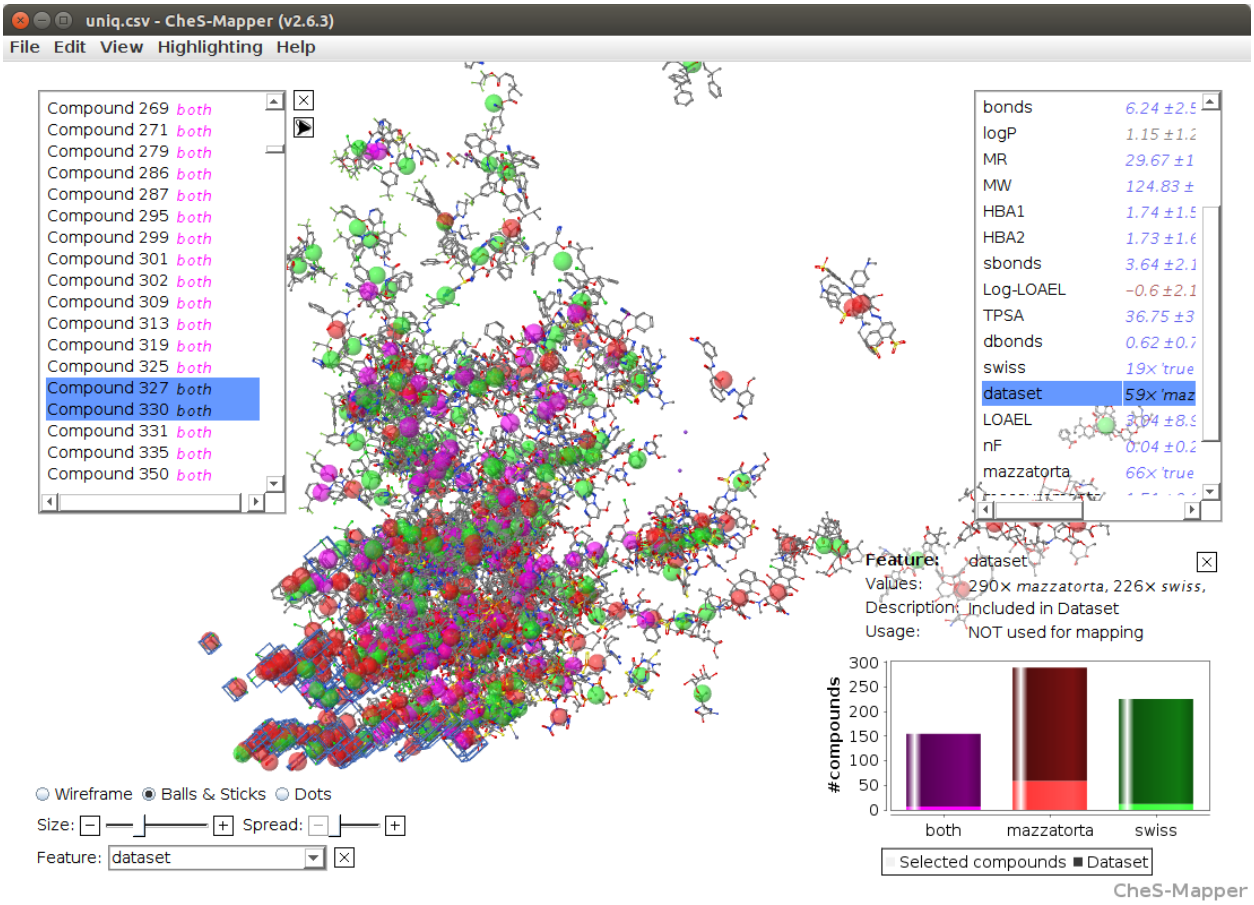

Figure 1: Compounds from the Mazzatorta and the Swiss dataset are highlighted in red and green. Compounds that occur in both datasets are highlighted in magenta. In this example, CheS-Mapper applied a principal components analysis to map compounds according to their physico-chemical (PC) feature values into 3D space. Both datasets have in general similar PC feature values. As an exception, the Mazzatorta dataset includes most of the tiny compound structures: we have selected the 78 smallest compounds (with 10 atoms and less, marked with a blue box in the screen-shot) and found that 61 of these compounds occur in the Mazzatorta dataset, whereas only 19 are contained in the Swiss dataset (p-value 3.7E-7).

Again, a visual inspection confirmed that both datasets are structurally very similar. However, CheS-Mapper allows the detection of features that help to distinguish groups of selected compounds from the entire dataset. Hence, we found discriminating features for compounds that occur in only one of both datasets, and for the most active or in-active compounds (see Table **??**). As an example, Figure 2 shows 9 compounds that match a specific fragment (all other compounds in the dataset do not match this fragment) and have very low mean LOAEL values.

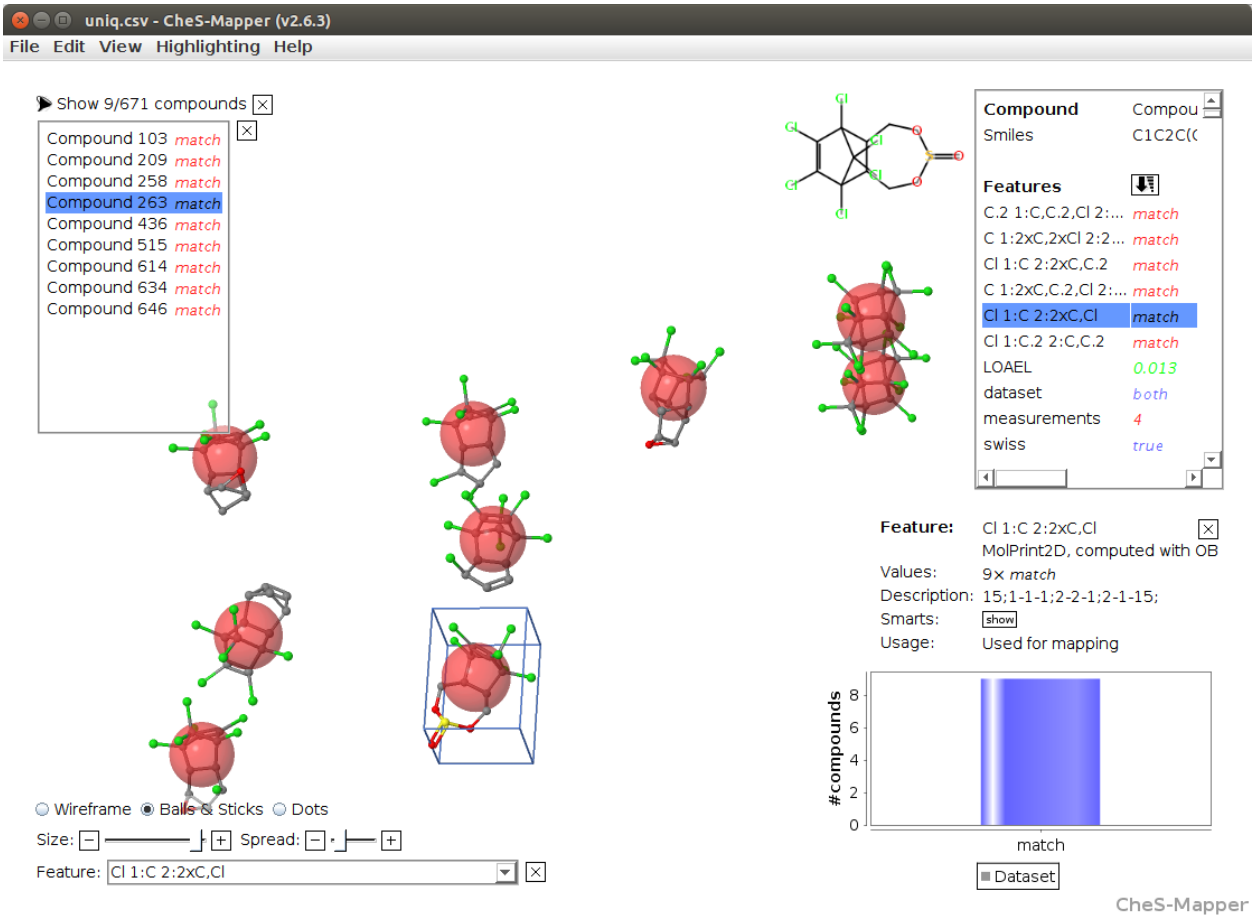

Figure 2: A CheS-Mapper screen-shot showing 9 compounds that match the MolPrint2D fragment 15;1-1-1;2-2-1;2-1-15; (as SMILES syntax: ClC(C)Cl). Apart from the selected compound (blue box), the other 8 compounds belong to the top 10 percent of compounds with the lowest LOAEL values. I.e., this feature can be regarded as a structural alert in our dataset, as it is matched by only 9 compounds in the entire dataset and 8 of these compounds are highly active.

Distribution of functional groups

In order to confirm the results of CheS-Mapper analysis we have evaluated the frequency of functional groups from the OpenBabel FP4 fingerprint. Figure 3 shows the frequency of functional groups in both datasets. Only 139 functional groups with a frequency > 25 are depicted, the complete table for all functional groups can be found in the data directory of the supplemental material (`data/functional-groups.csv`).

Experimental variability versus prediction uncertainty

Duplicated LOAEL values can be found in both datasets and there is a substantial number of 155 compounds occurring in both datasets. These duplicates allow us to estimate the variability of experimental results within individual datasets and between datasets.

Intra dataset variability

The Mazzatorta dataset has 567 LOAEL values for 445 unique structures, 93 compounds have multiple measurements with an average variance of 0.19 log10 units Figure 4.

The Swiss Federal Office dataset has 493 rat LOAEL values for 381 unique structures, 91 compounds have multiple measurements with a similar variance (average 0.15 log10 units). Variances of both datasets do not show a statistically significant difference with a p-value (t-test) of 0.25.

Inter dataset variability

Figure 5 shows the experimental LOAEL variability of compounds occurring in both datasets (i.e. the *test* dataset) colored in red (experimental). This is the baseline reference for the comparison with predicted values.

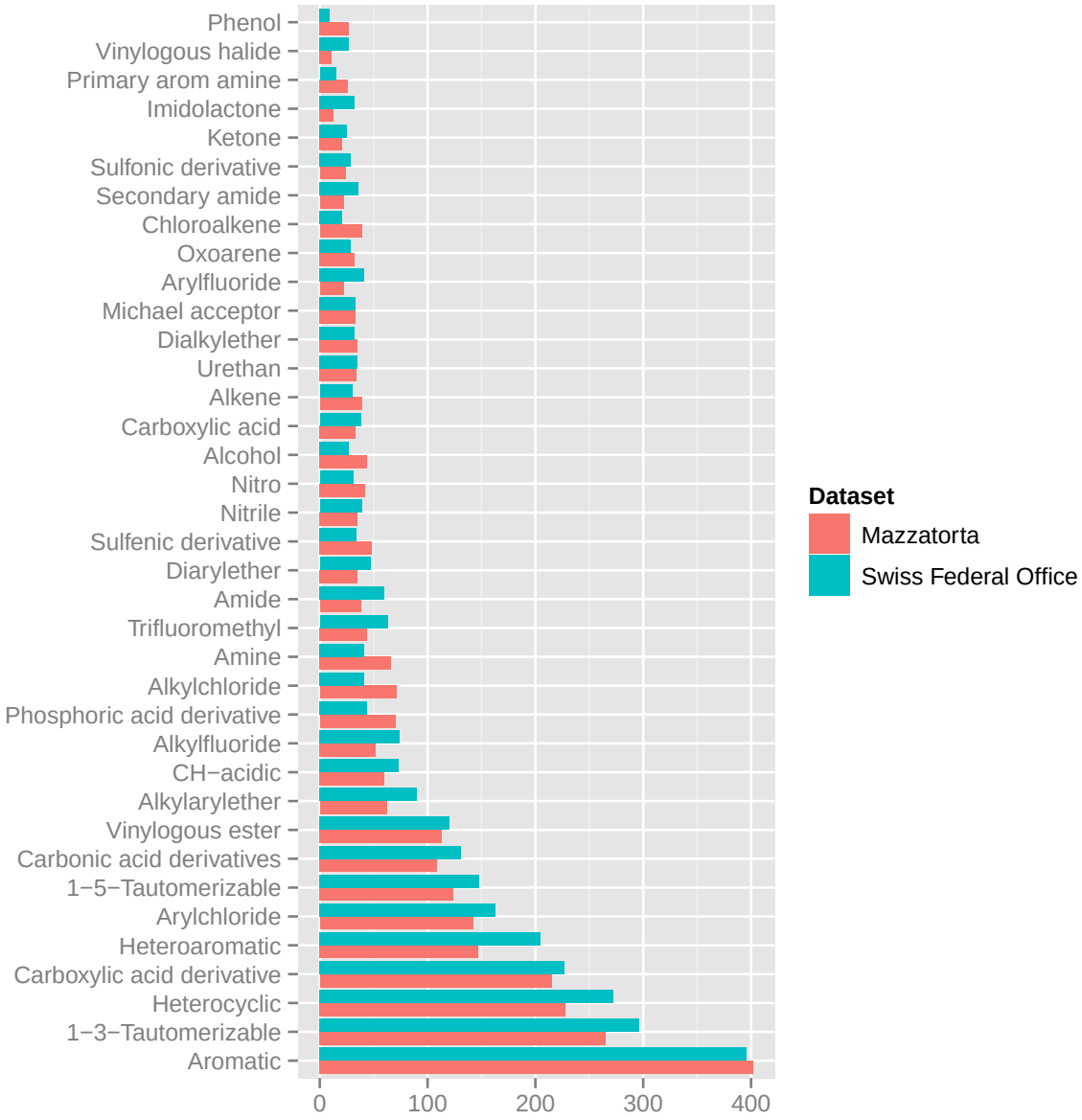

Figure 3: Frequency of functional groups.

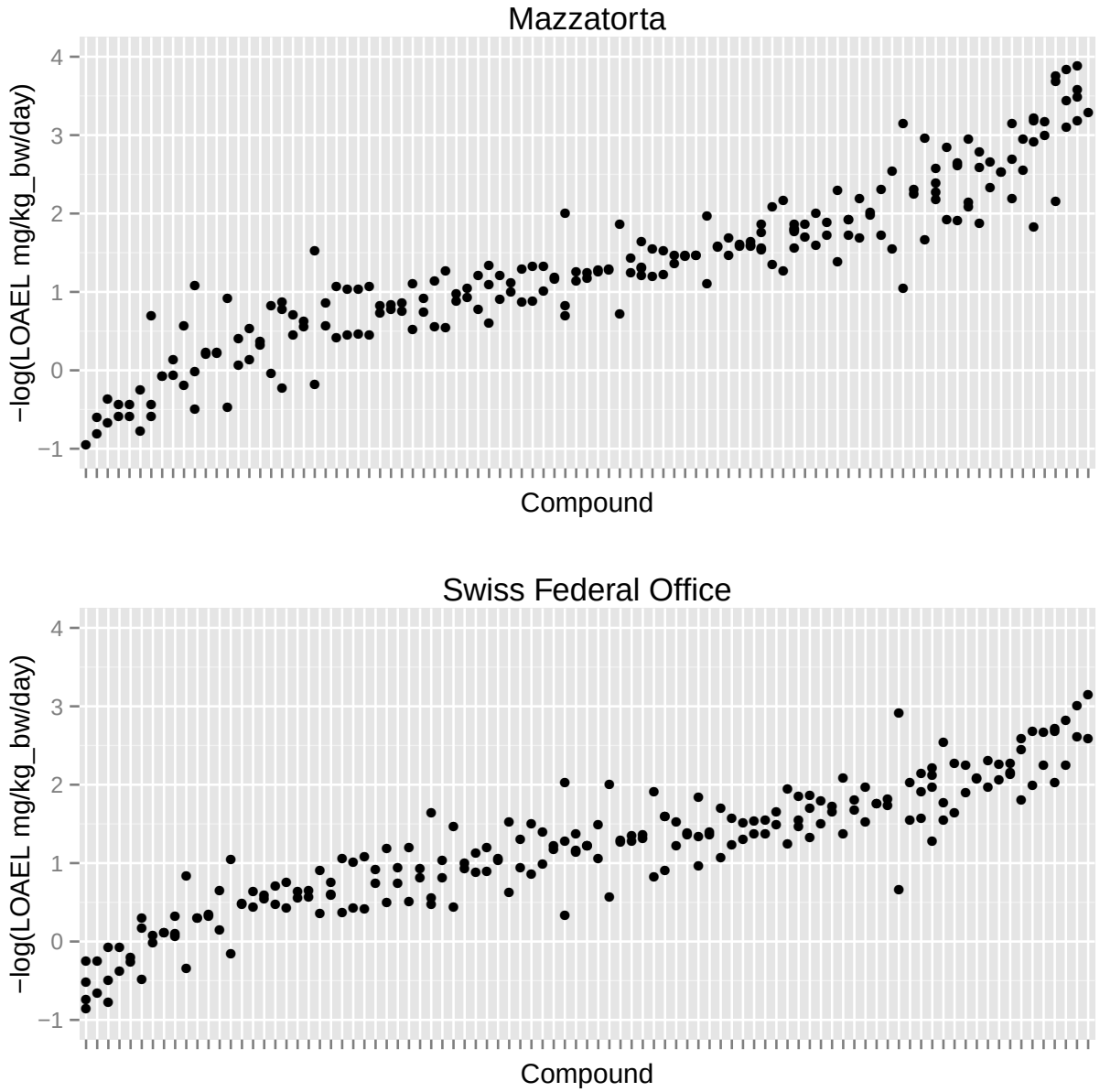

Figure 4: Distribution and variability of LOAEL values in both datasets: Each vertical line represents a compound, dots are individual LOAEL values.

LOAEL correlation between datasets

Figure 6 depicts the correlation between LOAEL values from both datasets. As both datasets contain duplicates we are using medians for the correlation plot and statistics. Please note that the aggregation of duplicated measurements into a single value hides a substantial portion of the real experimental variability. Correlation analysis shows a significant (p-value $<$ 2.2e-16) correlation between the experimental data in both datasets with r^2: 0.49, RMSE: 1.41

Local QSAR models

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 (375 measurements, 155 compounds).

The Mazzatorta, the Swiss Federal Office dataset and a combined dataset were used as training data for building `lazar` read across models. Predictions for the test set compounds were made after eliminating all information from the test compound from the corresponding training dataset. Figure 5 summarizes the results:

TODO: nr unpredicted, nr predictions outside of experimental values

Correlation analysis has been perfomed between individual predictions and the median of exprimental data. All correlations are statistically highly significant with a p-value $<$ 2.2e-16. These results are presented in Figure 6 and Table 2. Please bear in mind that the aggregation of experimental data into a single value actually hides experimental variability.

Table 1: Comparison of model predictions with experimental variability.

Training data	r^2	RMSE
Experimental	0.49	1.41

Training data	r^2	RMSE
Combined	0.41	1.47

TODO: repeated CV

Traditional 10-fold cross-validation results are summarised in Table 2 and Figure 7. All correlations are statistically highly significant with a p-value < 2.2e-16.

Table 2: 10-fold crossvalidation results

Training dataset	r^2	RMSE
Combined	0.39	1.84

Discussion

Elena + Benoit

- both datasets are structurally similar
- LOAEL values have similar variability in both datasets
- the Mazzatorta dataset has a small portion of very toxic compounds (low LOAEL, high -log10(LOAEL))
- lazar read across predictions fall within the experimental variability of LOAEL values
- predictions are slightly less accurate at extreme (high/low) LOAEL values, this can be explained by the algorithms used
- the original Mazzatorta paper has "better" results (R^2 0.54, RMSE 0.7) , but the model is likely to be overfitted (using genetic algorithms for feature selection *prior* to crossvalidation must lead to overfitted models)
- beware of over-optimisations and the race for "better" validation results

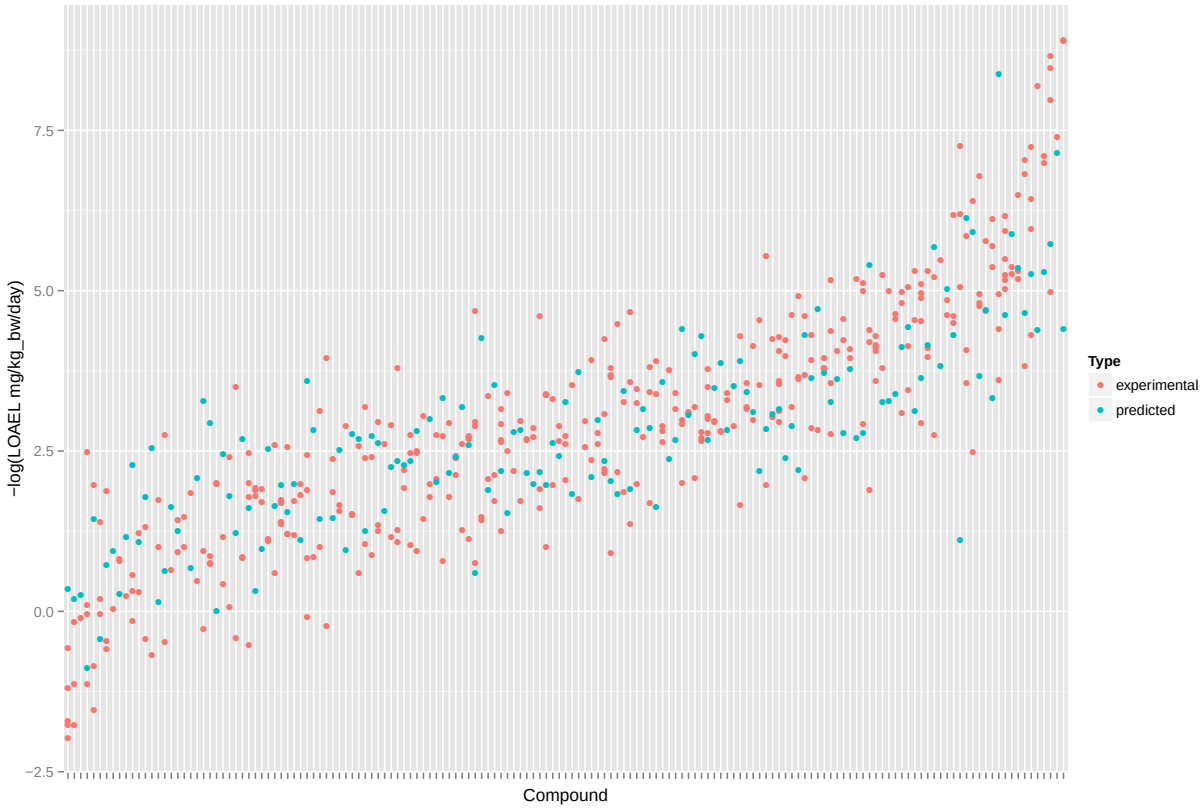

Figure 5: Comparison of experimental with predicted LOAEL values, each vertical line represents a compound, dots are individual measurements (red) or predictions (green).

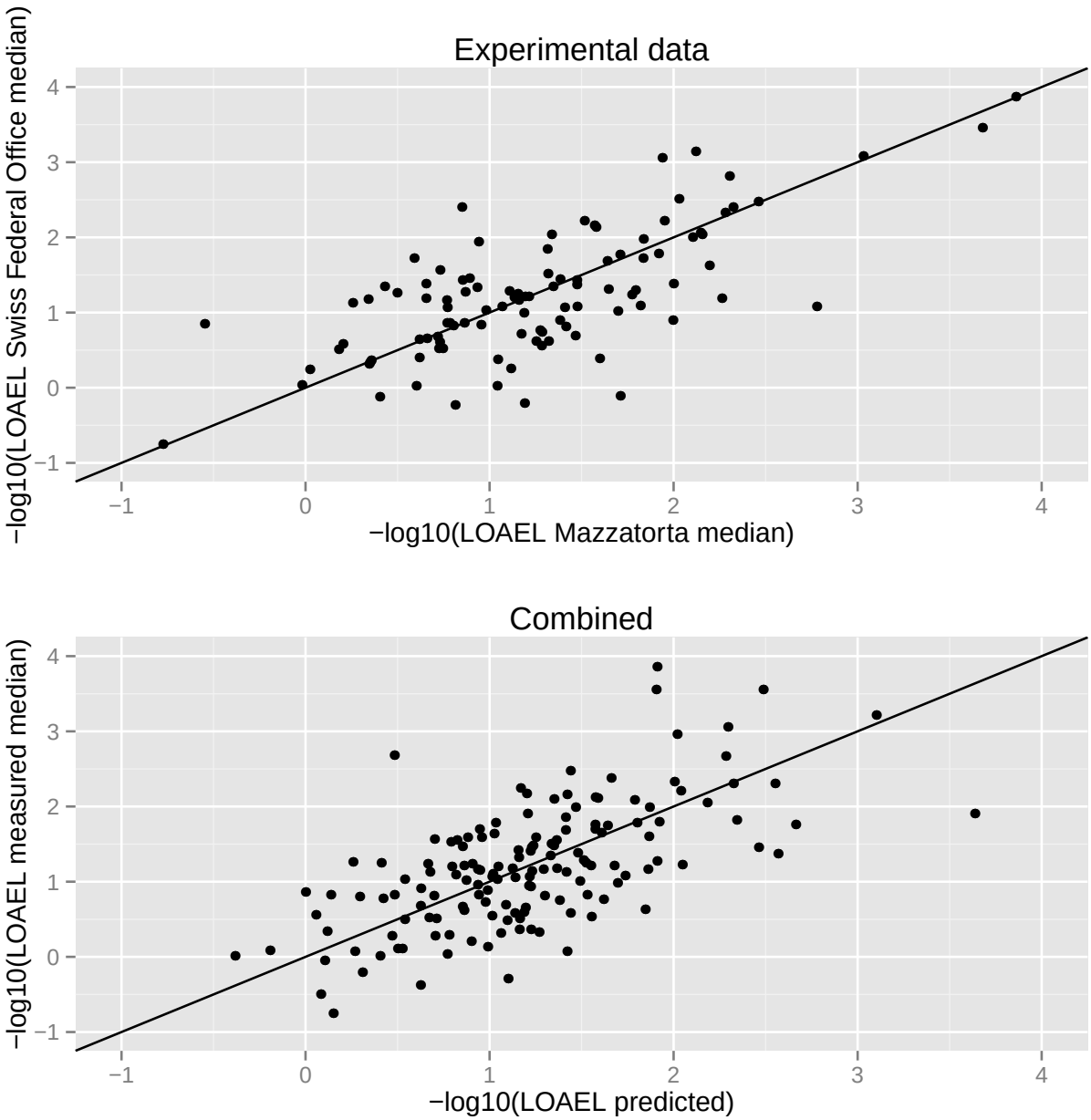

Figure 6: Correlation of experimental with predicted LOAEL values (test set)

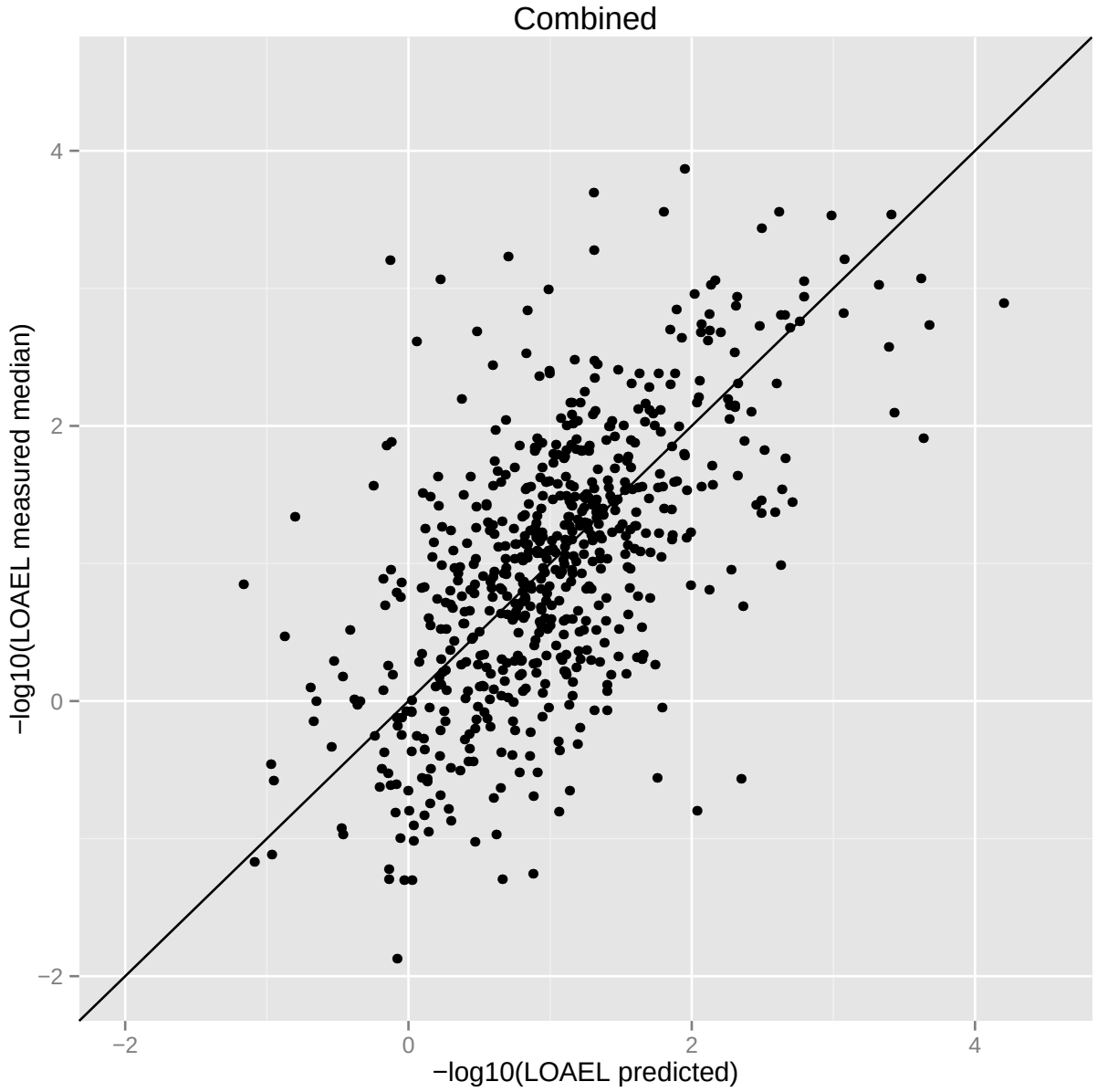

Figure 7: Correlation of experimental with predicted LOAEL values (10-fold crossvalidation)

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.

Kuhn, Max. 2008. "Building Predictive Models in R Using the Caret Package." *J. of Stat. Soft.*

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.

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.