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 usually not available.

With missing information about the variability of experimental toxicity data it is hard to judge the performance of predictive models objectively and it is tempting for model developers 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 the Mazzatorta and Swiss Federal Office datasets we will

- compare the structural diversity of both datasets
- compare the LOAEL values in both datasets
- build prediction models
- predict LOAELs of the test 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 GPL3 licenses.

A self-contained docker image with all programs, libraries and data required for the reproduction of these results is available from https://hub.docker.com/r/insilicotox/loael-paper/.

Source code and datasets for the reproduction of this manuscript can be downloaded from the GitHub repository https://github.com/opentox/loael-paper. The lazar framework (Maunz et al. 2013) is also available under a GPL3 License from https://github.com/opentox/lazar.

A graphical webinterface for `lazar` model predictions and validation results is publicly accessible at https://lazar.in-silico.ch, models presented in this manuscript will be included in future versions. Source code for the GUI can be obtained from https://github.com/opentox/lazar-gui.

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

Materials and Methods

The following sections give a high level overview about algorithms and datasets used for this study. In order to provide unambiguous references to algorithms and datasets, links to source code and data sources are included in the text.

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. The Mazzatorta dataset can be obtained from the following GitHub links: original data, unique smiles, -log10 transfomed LOAEL.

Swiss Federal Office dataset

Elena + Swiss Federal Office contribution (input)

The original Swiss Federal Office dataset has chronic toxicity data for rats, mice and multi generation effects. For the purpose of this study only rat LOAEL data with oral administration was used. This leads to the *Swiss Federal Office* dataset with 493 rat LOAEL values for 381 unique chemical structures. The Swiss dataset can be obtained from the following GitHub links: original data, unique smiles and mmol/kg_bw/day units, -log10 transfomed LOAEL.

Preprocessing

Chemical structures (represented as SMILES (Weininger 1988)) in both datasets were checked for correctness. 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 and rounded to five significant digits. For prediction, validation and visualisation purposes -log10 transformations are used.

Derived datasets

Two derived datasets were obtained from the original datasets:

The *test* dataset contains data from compounds that occur in both datasets. LOAEL values equal at five significant digits were considered as duplicates originating from the same study/publication and only one instance was kept in the test dataset. The test dataset has 375 LOAEL values for 155 unique chemical structures and was used for

- evaluating experimental variability
- comparing model predictions with experimental variaility.

The *training* dataset is the union of the Mazzatorta and the Swiss Federal Office dataset and it is used to build predictive models. LOAEL duplicates were removed using the same criteria as for the test dataset. The training 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. The complete `lazar` source code can be found on GitHub.

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. Unpublished 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).

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

The threshold selection is a trade-off between prediction accuracy (high threshold) and the number of predictable compounds (low threshold). As it is in many practical cases desirable to make predictions even in the absence of closely related neighbors, we follow a tiered approach:

First a similarity threshold of 0.5 is used to collect neighbors, to create a local QSAR model and to make a prediction for the query compound. If any of this steps fail, the procedure is repeated with a similarity threshold of 0.2 and the prediction is flagged with a warning that it might be out of the applicability domain of the training data.

Compounds with the same structure as the query structure are automatically eliminated from neighbors to obtain unbiased predictions in the presence of duplicates.

Local QSAR models and predictions

Only similar compounds (*neighbors*) above the threshold are used for local QSAR models. In this investigation we are using weighted random forests regression (RF) for the prediction of quantitative properties. First all uninformative fingerprints (i.e. features with identical values across all neighbors) are removed. The remaining set of features is used as descriptors for creating a local weighted RF model with atom environments as descriptors and model similarities as weights. The RF method from the `caret` R package (Kuhn 2008) is used for this purpose. Models are trained with the default `caret` settings, optimizing the number of RF components by bootstrap resampling.

Finally the local RF model is applied to predict the activity of the query compound. The RMSE of bootstrapped local model predictions is used to construct 95% prediction intervals at 1.96*RMSE.

If RF 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. In this case the prediction is also flagged with a warning.

Applicability domain

The applicability domain (AD) of lazar models is determined by the structural diversity of the training data. If no similar compounds are found in the training data no predictions will be generated. Warnings are issued if the similarity threshold has to be lowered from 0.5 to 0.2 in order to enable predictions and if lazar has to resort to weighted average predictions, because local random forests fail. Thus predictions without warnings can be considered as close to the applicability domain and predictions with warnings as more distant from the applicability domain. Quantitative applicability domain information can be obtained from the similarities of individual neighbors.

Local regression models consider neighbor similarities to the query compound, by weighting the contribution of each neighbor is by its similarity. The variability of local model predictions is reflected in the 95% prediction interval associated with each prediction.

Validation

For the comparison of experimental variability with predictive accuracies we are using a test set of compounds that occur in both datasets. Unbiased read across predictions are obtained from the *training* dataset, by removing *all* information from the test compound from the training set prior to predictions. This procedure is hardcoded into the prediction algorithm in order to prevent validation errors. As we have only a single test set no model or parameter optimisations were performed in order to avoid overfitting a single dataset.

Results from 3 repeated 10-fold crossvalidations with independent training/test set splits are provided as additional information to the test set results.

The final model for production purposes was trained with all available LOAEL data (Mazzatorta and Swiss Federal Office datasets combined).

Availability

Public webinterface https://lazar.in-silico.ch

`lazar` framework https://github.com/opentox/lazar (source code)

`lazar` GUI https://github.com/opentox/lazar-gui (source code)

Manuscript https://github.com/opentox/loael-paper (source code for the manuscript and validation experiments)

Docker image https://hub.docker.com/r/insilicotox/loael-paper/ (container with manuscript, validation experiments, `lazar` libraries and third party dependencies)

Results

Dataset comparison

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.

Structural diversity

In order to compare the structural diversity of both datasets we have evaluated the frequency of functional groups from the OpenBabel FP4 fingerprint. Figure 1 shows the frequency of functional groups in both datasets. 139 functional groups with a frequency > 25 are depicted, the complete table for all functional groups can be found in the supplemental material at GitHub.

This result was confirmed with a visual inspection using the CheS-Mapper (Chemical Space Mapping and Visualization in 3D, Gütlein, Karwath, and Kramer (2012)) tool. 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 depicts closely related (similar) compounds in 3D space and can be used with different kinds of features. We have investigated structural as well as physico-chemical properties and concluded that both datasets are very similar, both in terms of chemical structures and physico-chemical properties.

The only statistically significant difference between both datasets, is that the Mazzatorta dataset contains more small compounds (61 structures with less than 11 atoms) than the Swiss dataset (19 small structures, p-value 3.7E-7).

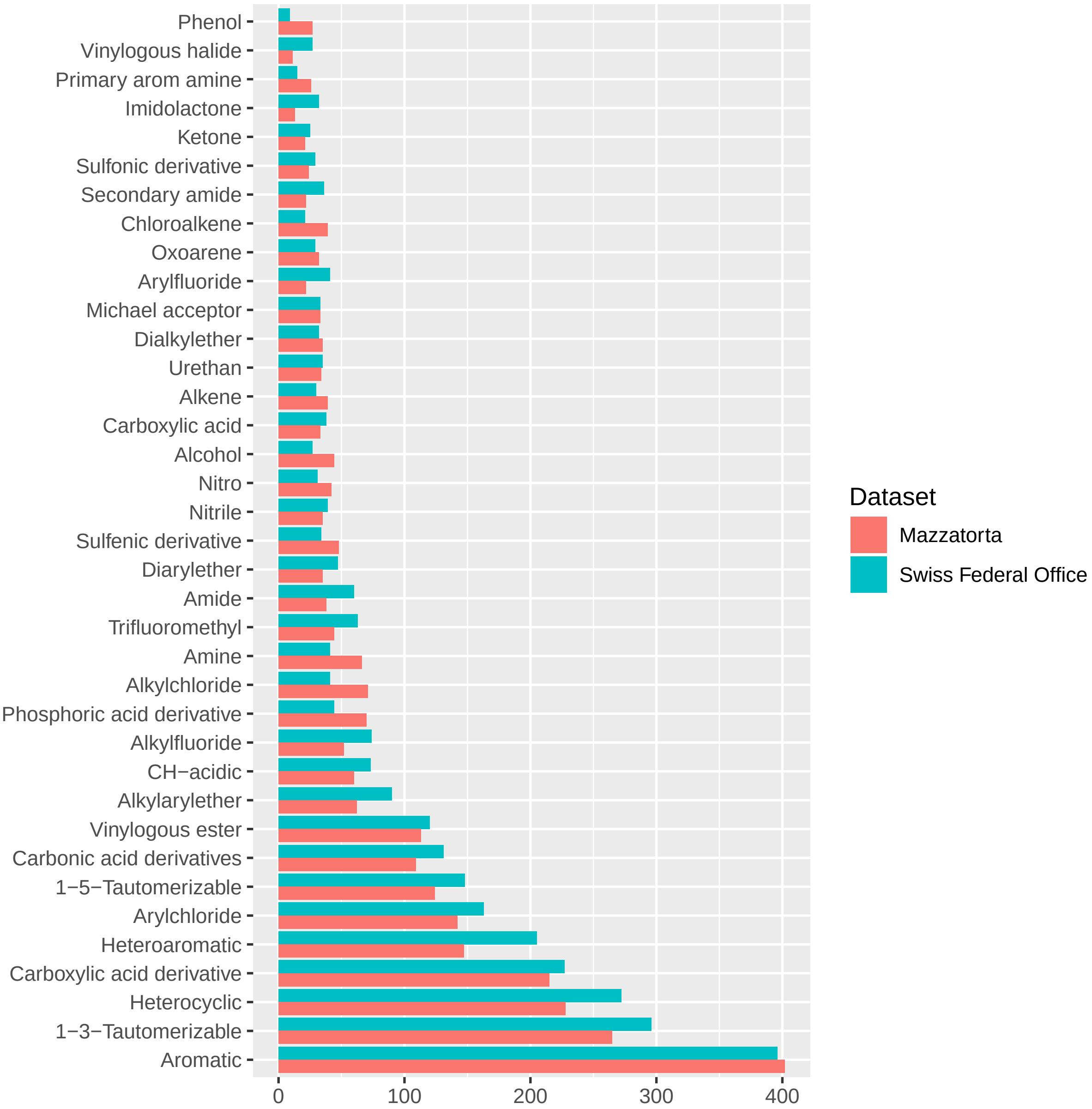


Figure 1: Frequency of functional groups.

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. Data with *identical* values (at five significant digits) in both datasets were excluded from variability analysis, because it it likely that they originate from the same experiments.

Intra dataset variability

The Mazzatorta dataset has 567 LOAEL values for 445 unique structures, 93 compounds have multiple measurements with a mean standard deviation (-log10 transformed values) of 0.32 (0.56 mg/kg_bw/day, 0.56 mmol/kg_bw/day) (Mazzatorta et al. (2008), Figure 2).

The Swiss Federal Office dataset has 493 rat LOAEL values for 381 unique structures, 91 compounds have multiple measurements with a mean standard deviation (-log10 transformed values) of 0.29 (0.57 mg/kg_bw/day, 0.59 mmol/kg_bw/day) (Figure 2).

Standard deviations of both datasets do not show a statistically significant difference with a p-value (t-test) of 0.21. The combined test set has a mean standard deviation (-log10 transformed values) of 0.33 (0.56 mg/kg_bw/day, 0.55 mmol/kg_bw/day) (Figure 2).

Inter dataset variability

Figure 4 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 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 median value hides a substantial

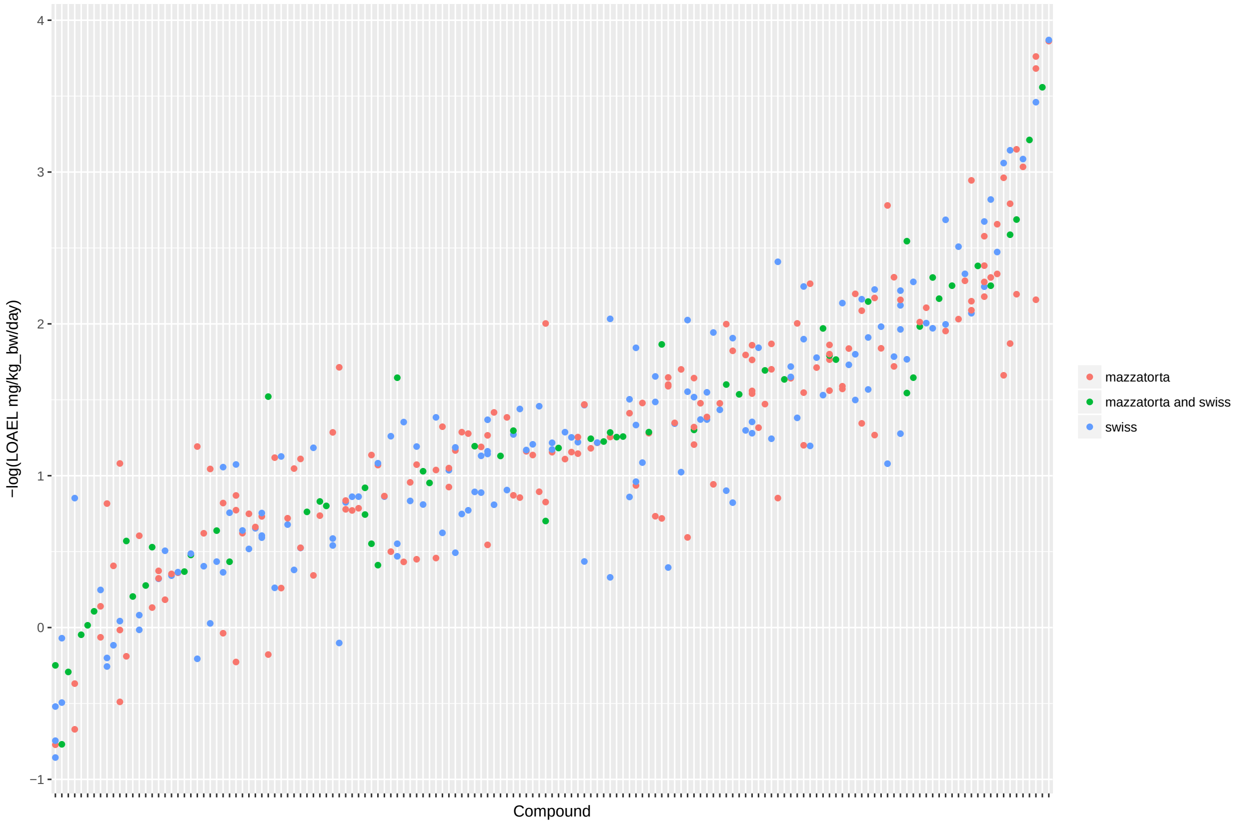


Figure 2: Distribution and variability of LOAEL values in both datasets. Each vertical line represents a compound, dots are individual LOAEL values.

portion of the experimental variability. Correlation analysis shows a significant (p-value < 2.2e-16) correlation between the experimental data in both datasets with rˆ2: 0.52, RMSE: 0.59

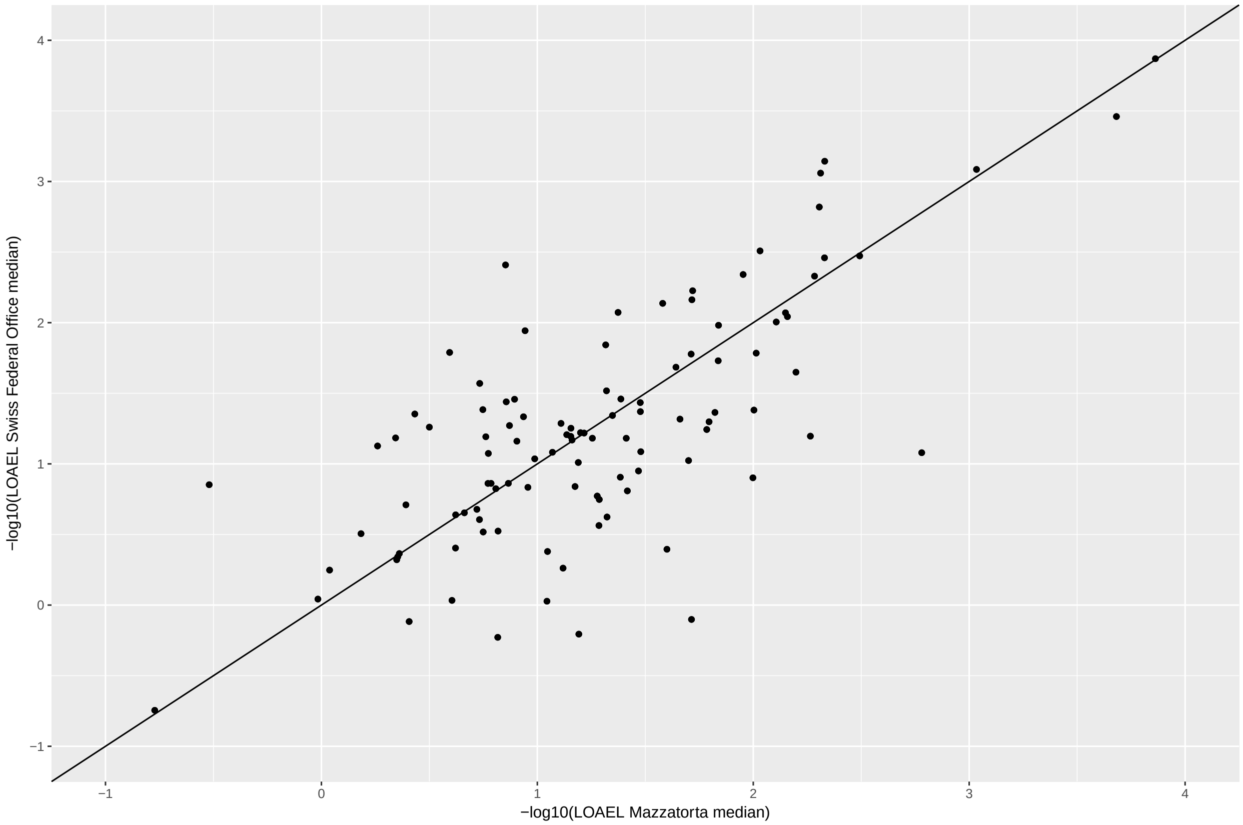


Figure 3: Correlation of median LOAEL values from Mazzatorta and Swiss datasets. Data with identical values in both datasets was removed from analysis.

Local QSAR models

In order to compare the performance of in silico read across models with experimental variability we are using compounds that occur in both datasets as a test set (375 measurements, 155 compounds). `lazar` read across predictions were obtained for 155 compounds, 37 predictions failed, because no similar compounds were found in the training data (i.e. they were not covered by the applicability domain of the training data).

Experimental data and 95% prediction intervals overlapped in 100% of the test examples.

Figure 4 shows a comparison of predicted with experimental values:

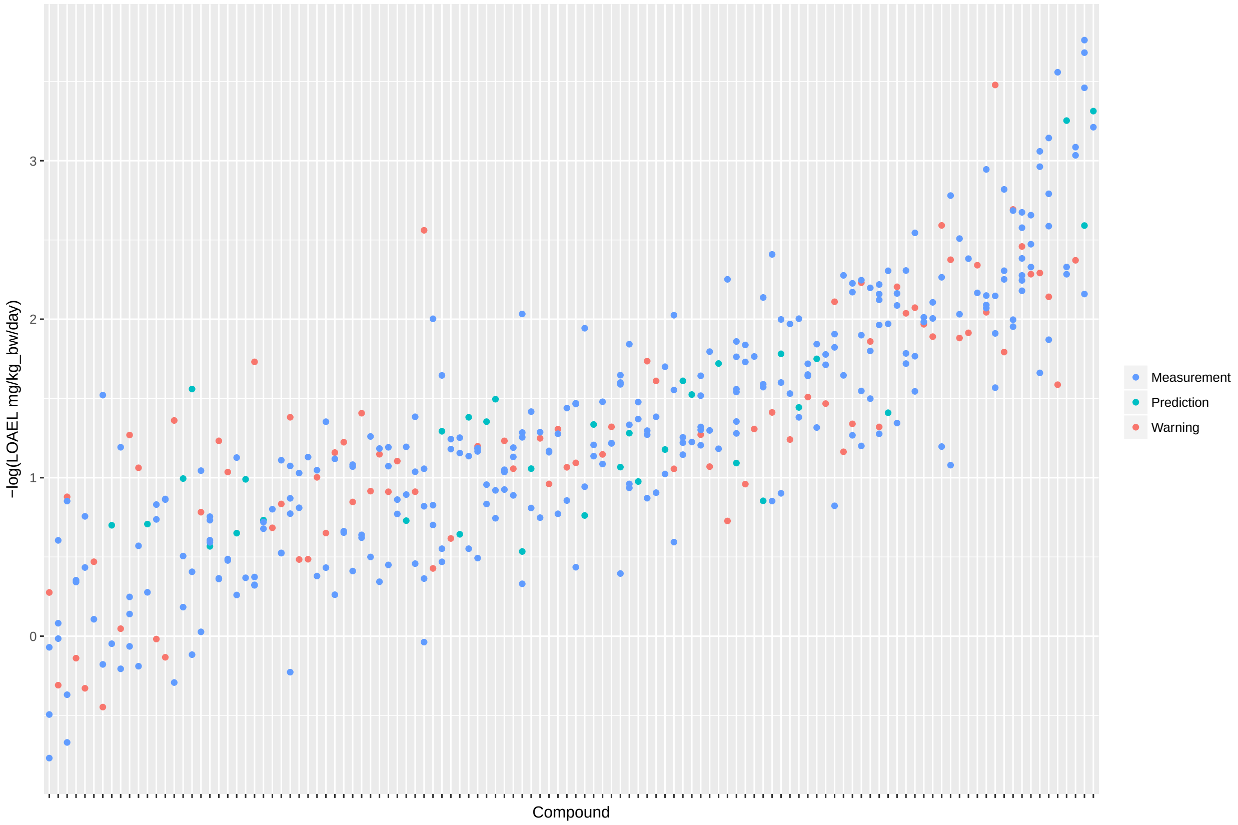


Figure 4: Comparison of experimental with predicted LOAEL values. Each vertical line represents a compound, dots are individual measurements (blue), predictions (green) or predictions far from the applicability domain, i.e. with warnings (red).

Correlation analysis was performed between individual predictions and the median of experimental data. All correlations are statistically highly significant with a p-value < 2.2e-16. These results are presented in Figure 5 and Table 2. Please bear in mind that the aggregation of multiple measurements into a single median value hides experimental variability.

Table 1: Comparison of model predictions with experimental variability.

Comparison	r^2	RMSE	Nr. predicted
Mazzatorta vs. Swiss dataset	0.52	0.59	
AD close predictions vs. test median	0.48	0.56	34/155
AD distant predictions vs. test median	0.38	0.68	84/155

Comparison	r^2	RMSE	Nr. predicted
All predictions vs. test median	0.4	0.65	118/155

For a further assessment of model performance three independent 10-fold cross-validations were performed. Results are summarised in Table 2 and Figure 6. All correlations of predicted with experimental values are statistically highly significant with a p-value $< 2.2e\text{-}16$.

Table 2: Results from 3 independent 10-fold crossvalidations

Predictions	r^2	RMSE	Nr. predicted
AD close	0.61	0.58	102/671
AD distant	0.45	0.78	374/671
All	0.47	0.74	476/671
AD close	0.59	0.6	101/671
AD distant	0.45	0.77	376/671
All	0.47	0.74	477/671
AD close	0.59	0.57	93/671
AD distant	0.43	0.81	384/671
All	0.45	0.77	477/671

Discussion

Elena + Benoit

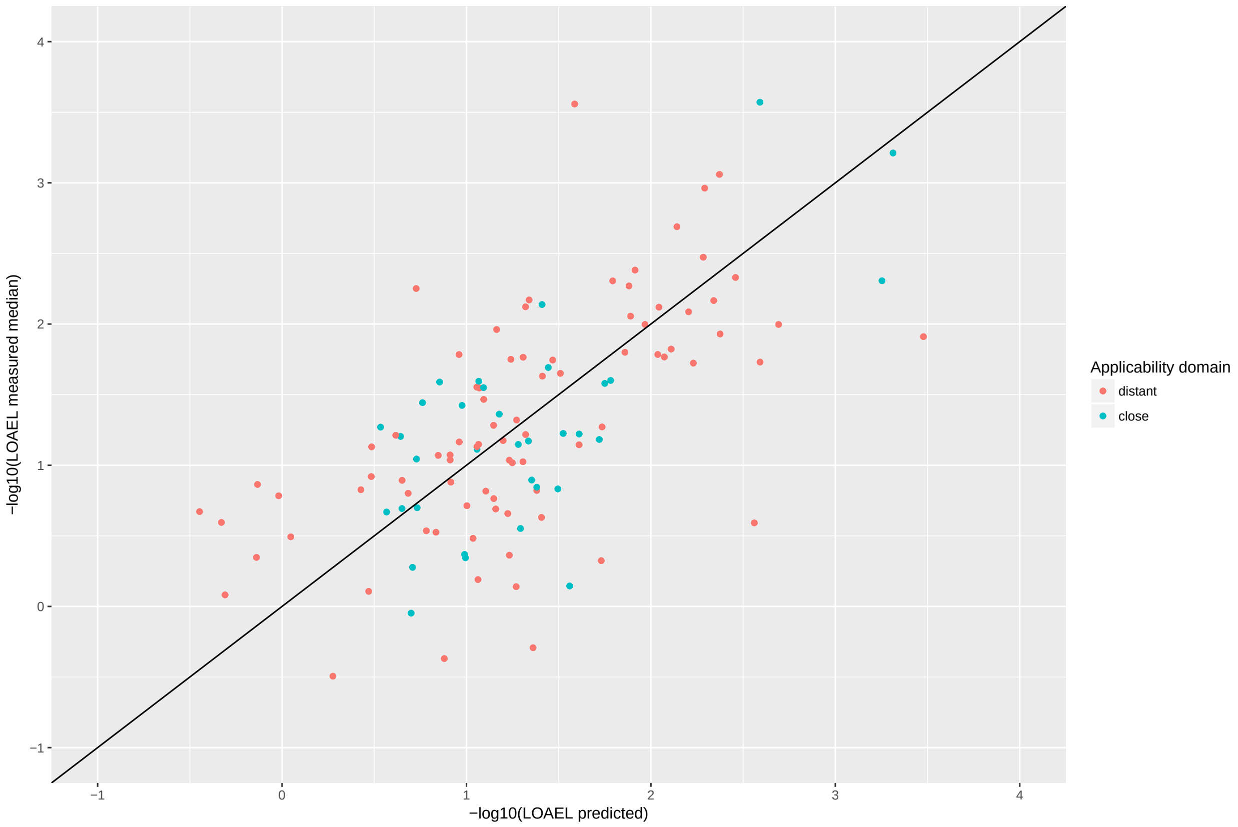
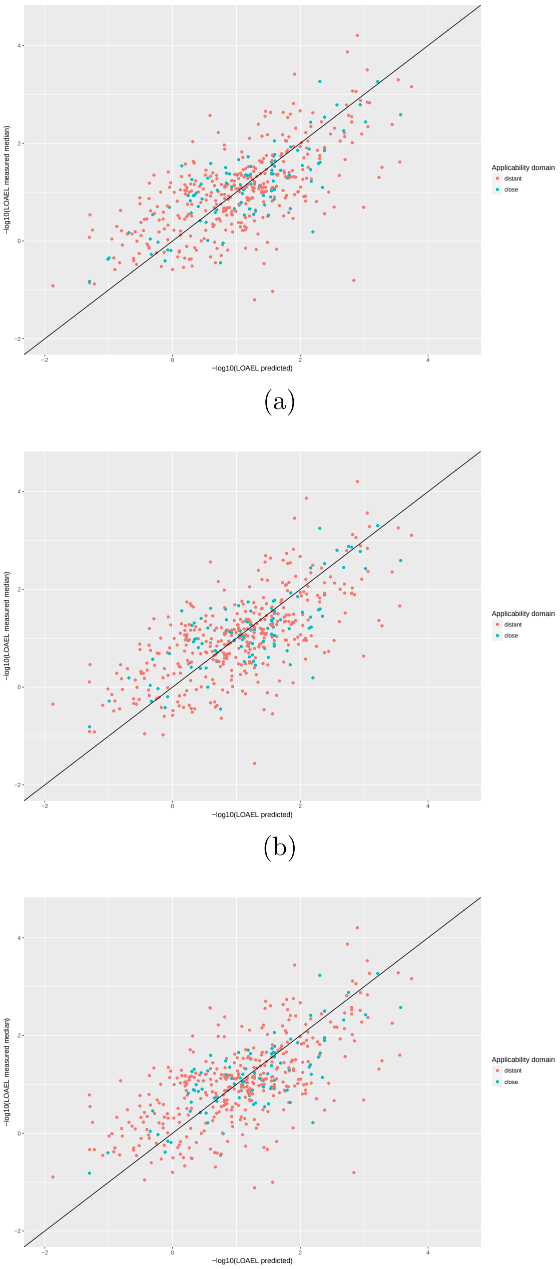


Figure 5: Correlation of experimental with predicted LOAEL values (test set). Green dots indicate predictions close to the applicability domain (i.e. without warnings), red dots indicate predictions far from the applicability domain (i.e. with warnings).



(a)

(b)

(c)

Figure 6: Correlation of predicted vs. measured values for three independent crossvalidations with *MP2D* fingerprint descriptors and local *random forest* models

Dataset comparison

Our investigations clearly indicate that the Mazzatorta and Swiss Federal Office datasets are very similar in terms of chemical structures and properties and the distribution of experimental LOAEL values. The only significant difference that we have observed was that the Mazzatorta dataset has larger amount of small molecules, than the Swiss Federal Office dataset. For this reason we have pooled both dataset into a single training dataset for read across predictions.

Figure 2 and Figure 5 and Table 1 show however considerable variability in the experimental data. High experimental variability has an impact on model building and on model validation. First it influences model quality by introducing noise into the training data, secondly it influences accuracy estimates because predictions have to be compared against noisy data where "true" experimental values are unknown. This will become obvious in the next section, where we compare predictions with experimental data.

`lazar` predictions

Table 1, Table 2, Figure 4, Figure 5 and Figure 6 clearly indicate that `lazar` generates reliable predictions for compounds within the applicability domain of the training data (i.e. predictions without warnings, which indicates a sufficient number of neighbors with similarity > 0.5 to create local random forest models). Correlation analysis (Table 1, Table 2) shows, that errors ($RMSE$) and explained variance (r^2) are comparable to experimental variability of the training data.

Predictions with a warning (neighbor similarity < 0.5 and > 0.2 or weighted average predictions) are a grey zone. They still show a strong correlation with experimental data, but the errors are larger than for compounds within the applicability domain (Table 1, Table 2). Expected errors are displayed as 95% prediction intervals, which covers 100% of the experimental data. The main advantage of lowering the similarity threshold is that it allows to predict a much larger number of substances than with more rigorous applicability domain criteria. As each of this prediction could be problematic, they are flagged with a warning to alert risk assessors that further inspection is required. This can be done in the graphical interface (https://lazar.in-silico.ch) which provides intuitive means of inspecting the rationales and data used for read across predictions.

Finally there is a substantial number of compounds (37), where no predictions can be made, because there are no similar compounds in the training data. These compounds clearly fall beyond the applicability domain of the training dataset and in such cases it is preferable to avoid predictions instead of random guessing.

Elena: Should we add a GUI screenshot?

Summary

We could demonstrate that `lazar` predictions within the applicability domain of the training data have the same variability as the experimental training data. In such cases experimental investigations can be substituted with in silico predictions. Predictions with a lower similarity threshold can still give usable results, but the errors to be expected are higher and a manual inspection of prediction results is highly recommended.

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.