

1

Modeling Chronic Toxicity: A comparison of experimental variability with (Q)SAR/read-across predictions

2
Christoph Helma¹ David Vorgrimmler¹ Denis Gebele¹ Martin Gütlein² Barbara
Engeli³ Jürg Zarn³ Benoit Schilter⁴ Elena Lo Piparo⁴

E-mail:

3
Abstract

4 This study compares the accuracy of (Q)SAR/read-across predictions with the
5 experimental variability of chronic lowest-observed-adverse-effect levels (LOAELs)
6 from *in vivo* experiments. We could demonstrate that predictions of the lazy structure-
7 activity relationships (lazar) algorithm within the applicability domain of the training
8 data have the same variability as the experimental training data. Predictions with a
9 lower similarity threshold (i.e. a larger distance from the applicability domain) are
10 also significantly better than random guessing, but the errors to be expected are
11 higher and a manual inspection of prediction results is highly recommended.

12 ¹ in silico toxicology gmbh, Basel, Switzerland

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

14 ³ Federal Food Safety and Veterinary Office (FSVO) , Risk Assessment Division , Bern ,
15 Switzerland

16 ⁴ Chemical Food Safety Group, Nestlé Research Center, Lausanne, Switzerland

17 **Introduction**

18 Relying on standard animal toxicological testing for chemical hazard identification and
19 characterization is increasingly questioned on both scientific and ethical grounds. In
20 addition, it appears obvious that from a resource perspective, the capacity of standard
21 toxicology to address the safety of thousands of untested chemicals (Fowler, Savage, and
22 Mendez 2011) to which human may be exposed is very limited. It has also been recognized
23 that getting rapid insight on toxicity of chemicals in case of emergency safety incidents or
24 for early prioritization in research and development (safety by design) is a big challenge
25 mainly because of the time and cost constraints associated with the generation of relevant
26 animal data. In this context, alternative approaches to obtain timely and fit-for-purpose
27 toxicological information are being developed. Amongst others *in silico* toxicology methods
28 are considered highly promising. Importantly, they are raising more and more interests
29 and getting increased acceptance in various regulatory (e.g. (ECHA 2008, EFSA (2016),
30 EFSA (2014), Health Canada (2016), OECD (2015))) and industrial (e.g. (Stanton and
31 Krusezewski 2016, Lo Piparo et al. (2011))) frameworks.

32 For a long time already, computational methods have been an integral part of pharmaceutical
33 discovery pipelines, while in chemical food safety their actual potentials emerged only
34 recently (Lo Piparo et al. 2011). In this field, an application considered critical is in
35 the establishment of levels of safety concern in order to rapidly and efficiently manage
36 toxicologically uncharacterized chemicals identified in food. This requires a risk-based
37 approach to benchmark exposure with a quantitative value of toxicity relevant for risk
38 assessment (Schilter et al. 2014). Since chronic studies have the highest power (more animals

39 per group and more endpoints than other studies) and because long-term toxicity studies are
40 often the most sensitive in food toxicology databases, predicting chronic toxicity is of prime
41 importance. Up to now, read-across and Quantitative Structure Activity Relationships
42 (QSAR) have been the most used *in silico* approaches to obtain quantitative predictions of
43 chronic toxicity.

44 The quality and reproducibility of (Q)SAR and read-across predictions has been a con-
45 tinuous and controversial topic in the toxicological risk-assessment community. Although
46 model predictions can be validated with various procedures, to review results in context
47 of experimental variability has actually been rarely done or attempted. With missing
48 information about the variability of experimental toxicity data it is hard to judge the
49 performance of predictive models objectively and it is tempting for model developers to use
50 aggressive model optimisation methods that lead to impressive validation results, but also
51 to overfitted models with little practical relevance.

52 In the present study, automatic read-across like models were built to generate quantitative
53 predictions of long-term toxicity. The aim of the work was not to predict the nature of
54 the toxicological effects of chemicals, but to obtain quantitative values which could be
55 compared to exposure. Two databases compiling chronic oral rat Lowest Adverse Effect
56 Levels (LOAEL) as endpoint were used. An early review of the databases revealed that many
57 chemicals had at least two independent studies/LOAELs. These studies were exploited
58 to generate information on the reproducibility of chronic animal studies and were used to
59 evaluate prediction performance of the models in the context of experimental variability.

60 An important limitation often raised for computational toxicology is the lack of transparency
61 on published models and consequently on the difficulty for the scientific community to
62 reproduce and apply them. To overcome these issues, source code for all programs and
63 libraries and the data that have been used to generate this manuscript are made available

64 under GPL3 licenses. Data and compiled programs with all dependencies for the reproduc-
65 tion of results in this manuscript are available as a self-contained docker image. All data,
66 tables and figures in this manuscript was generated directly from experimental results using
67 the R package `knitr`.

68 **Materials and Methods**

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

72 **Datasets**

73 **Nestlé database**

74 The first database (Nestlé database for further reference) originates from the publication of
75 (P. Mazzatorta et al. 2008). It contains chronic (> 180 days) lowest observed effect levels
76 (LOAEL) for rats (*Rattus norvegicus*) after oral (gavage, diet, drinking water) administration.
77 The Nestlé database consists of 567 LOAEL values for 445 unique chemical structures. The
78 Nestlé database can be obtained from the following GitHub links:

- 79 • original data: https://github.com/opentox/loael-paper/blob/revision/data/LOAEL_mg_corrected_smiles_mmol.csv
- 80 • unique smiles: <https://github.com/opentox/loael-paper/blob/revision/data/mazzatorta.csv>
- 81 • -log10 transformed LOAEL: https://github.com/opentox/loael-paper/blob/revision/data/mazzatorta_log10.csv.
- 82
- 83
- 84

85 Swiss Food Safety and Veterinary Office (FSVO) database

86 Publicly available data from pesticide evaluations of chronic rat toxicity studies from the
87 European Food Safety Authority (EFSA) (EFSA 2014), the Joint FAO/WHO Meeting on
88 Pesticide Residues (JMPR) (WHO 2011) and the US EPA (US EPA 2011) were compiled
89 to form the FSVO-database. Only studies providing both an experimental NOAEL and an
90 experimental LOAEL were included. The LOAELs were taken as they were reported in the
91 evaluations. Further details on the database are described elsewhere (Zarn, Engeli, and
92 Schlatter 2011, Zarn, Engeli, and Schlatter (2013)). The FSVO-database consists of 493 rat
93 LOAEL values for 381 unique chemical structures. It can be obtained from the following
94 GitHub links:

- 95 • original data: [https://github.com/opentox/loael-paper/blob/revision/data/](https://github.com/opentox/loael-paper/blob/revision/data/NOAEL-LOAEL_SMILES_rat_chron.csv)
96 [NOAEL-LOAEL_SMILES_rat_chron.csv](https://github.com/opentox/loael-paper/blob/revision/data/NOAEL-LOAEL_SMILES_rat_chron.csv)
- 97 • unique smiles and mmol/kg_bw/day units: [https://github.com/opentox/loael-paper/](https://github.com/opentox/loael-paper/blob/revision/data/swiss.csv)
98 [blob/revision/data/swiss.csv](https://github.com/opentox/loael-paper/blob/revision/data/swiss.csv)
- 99 • -log10 transformed LOAEL: [https://github.com/opentox/loael-paper/blob/revision/](https://github.com/opentox/loael-paper/blob/revision/data/swiss_log10.csv)
100 [data/swiss_log10.csv](https://github.com/opentox/loael-paper/blob/revision/data/swiss_log10.csv)

101 Preprocessing

102 Chemical structures (represented as SMILES (Weininger 1988)) in both databases were
103 checked for correctness. When syntactically incorrect or missing SMILES were generated
104 from other identifiers (e.g names, CAS numbers). Unique smiles from the OpenBabel library
105 (OBoyle et al. 2011) were used for the identification of duplicated structures.

106 Studies with undefined or empty LOAEL entries were removed from the databases. LOAEL
107 values were converted to mmol/kg bw/day units and rounded to five significant digits. For

108 prediction, validation and visualisation purposes $-\log_{10}$ transformations are used.

109 **Derived datasets**

110 Two derived datasets were obtained from the original databases:

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

- 115 • evaluating experimental variability
- 116 • comparing model predictions with experimental variability.

117 The *training dataset* is the union of the Nestlé and the FSVO databases and it was used
118 to build predictive models. LOAEL duplicates were removed using the same criteria as
119 for the test dataset. The training dataset has 998 LOAEL values for 671 unique chemical
120 structures.

121 **Algorithms**

122 In this study we are using the modular *lazar* (*lazy structure activity relationships*) framework
123 (A. Maunz et al. 2013) for model development and validation. The complete `lazar` source
124 code can be found on [GitHub](#).

125 `lazar` follows the following basic [workflow](#):

126 For a given chemical structure `lazar`

- 127 • searches in a database for similar structures (*neighbors*) with experimental data,
- 128 • builds a local QSAR model with these neighbors and

129 • uses this model to predict the unknown activity of the query compound.

130 This procedure resembles an automated version of *read across* predictions in toxicology, in
131 machine learning terms it would be classified as a *k-nearest-neighbor* algorithm.

132 Apart from this basic workflow *lazar* is completely modular and allows the researcher to use
133 any algorithm for similarity searches and local QSAR modelling. Algorithms used within
134 this study are described in the following sections.

135 **Neighbor identification**

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

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

141 MolPrint2D fingerprints are generated dynamically from chemical structures and do not
142 rely on predefined lists of fragments (such as OpenBabel FP3, FP4 or MACCs fingerprints
143 or lists of toxicophores/toxicophobes). This has the advantage that they may capture
144 substructures of toxicological relevance that are not included in other fingerprints.

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

147 The [chemical similarity](#) between two compounds A and B is expressed as the proportion
148 between atom environments common in both structures $A \cap B$ and the total number of
149 atom environments $A \cup B$ (Jaccard/Tanimoto index, Equation 1).

$$sim = \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

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

- 154 • First a similarity threshold of 0.5 is used to collect neighbors, to create a local QSAR
155 model and to make a prediction for the query compound.
- 156 • If any of these steps fails, the procedure is repeated with a similarity threshold of 0.2
157 and the prediction is flagged with a warning that it might be out of the applicability
158 domain of the training data.
- 159 • Similarity thresholds of 0.5 and 0.2 are the default values chosen by the software
160 developers and remained unchanged during the course of these experiments.

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

163 **Local QSAR models and predictions**

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

171 RF components by bootstrap resampling.

172 Finally the local RF model is applied to **predict the activity** of the query compound. The
173 root-mean-square error (RMSE) of bootstrapped local model predictions is used to construct
174 95% prediction intervals at $1.96 \times \text{RMSE}$. The width of the prediction interval indicates
175 the expected prediction accuracy. The “true” value of a prediction should be with 95%
176 probability within the prediction interval.

177 If RF modelling or prediction fails, the program resorts to using the **weighted mean** of
178 the neighbors LOAEL values, where the contribution of each neighbor is weighted by its
179 similarity to the query compound. In this case the prediction is also flagged with a warning.

180 **Applicability domain**

181 The applicability domain (AD) of lazar models is determined by the structural diversity of
182 the training data. If no similar compounds are found in the training data no predictions will
183 be generated. Warnings are issued if the similarity threshold has to be lowered from 0.5 to
184 0.2 in order to enable predictions and if lazar has to resort to weighted average predictions,
185 because local random forests fail. Thus predictions without warnings can be considered as
186 close to the applicability domain and predictions with warnings as more distant from the
187 applicability domain. Quantitative applicability domain information can be obtained from
188 the similarities of individual neighbors.

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

192 **Validation**

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

199 Results from 50 repeated 10-fold crossvalidations with independent training/test set splits
200 are provided as additional information to the test set results.

201 The final model for production purposes was trained with all available LOAEL data (Nestlé
202 and FSVO databases combined).

203 **Availability**

204 **Public webinterface** <https://lazar.in-silico.ch> (see Figure 1)

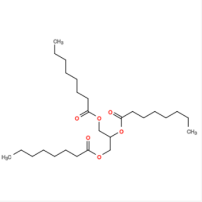
205 **lazar framework** <https://github.com/opentox/lazar> (source code)

206 **lazar GUI** <https://github.com/opentox/lazar-gui> (source code)

207 **Manuscript** <https://github.com/opentox/loael-paper/tree/revision> (source code for the
208 manuscript and validation experiments)

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

Prediction Results:

Compound	Lowest observed adverse effect level (LOAEL) (Rodents)
 <chem>CCCCCCCC(=O)OC(COC(=O)CCCCCCC)COC(=O)CCCCCCC</chem>	Prediction: 9.29 (mmol/kg_bw/day) 4370.0 (mg/kg_bw/day) 95% Prediction interval: 1.43 - 60.4 (mmol/kg_bw/day) 671.0 - 28400.0 (mg/kg_bw/day)

Neighbors:

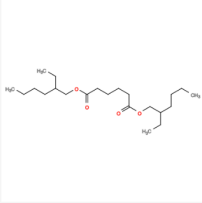
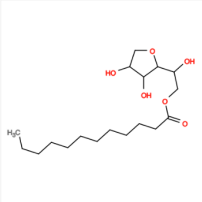
Lowest observed adverse effect level (LOAEL) (Rodents)		
Compound	Measured Activity i	Similarity i
 <chem>CCCC(COC(=O)CCCC(=O)OCC(CCCC)CC)OC</chem>	4.05 (mmol/kg_bw/day) 1500.0 (mg/kg_bw/day)	0.571
 <chem>CCCCCCCCCCCC(=O)OCC(C1OCC(C1O)O)O</chem>	19.9 (mmol/kg_bw/day) 6880.0 (mg/kg_bw/day)	0.529

Figure 1: Screenshot of a lazar prediction from the public webinterface.

211 Results

212 Dataset comparison

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

216 Structural diversity

217 In order to compare the structural diversity of both databases we evaluated the frequency
218 of functional groups from the OpenBabel FP4 fingerprint. Figure 2 shows the frequency
219 of functional groups in both databases. 139 functional groups with a frequency > 25 are
220 depicted, the complete table for all functional groups can be found in the supplemental
221 material at [GitHub](#).

222 This result was confirmed with a visual inspection using the [CheS-Mapper](#) (Chemical
223 Space Mapping and Visualization in 3D, Gütlein, Karwath, and Kramer (2012)) tool.
224 CheS-Mapper can be used to analyze the relationship between the structure of chemical
225 compounds, their physico-chemical properties, and biological or toxic effects. It depicts
226 closely related (similar) compounds in 3D space and can be used with different kinds
227 of features. We have investigated structural as well as physico-chemical properties and
228 concluded that both databases are very similar, both in terms of chemical structures and
229 physico-chemical properties.

230 The only statistically significant difference between both databases is that the Nestlé
231 database contains more small compounds (61 structures with less than 11 non-hydrogen
232 atoms) than the FSVO-database (19 small structures, chi-square test: p-value 3.7E-7).

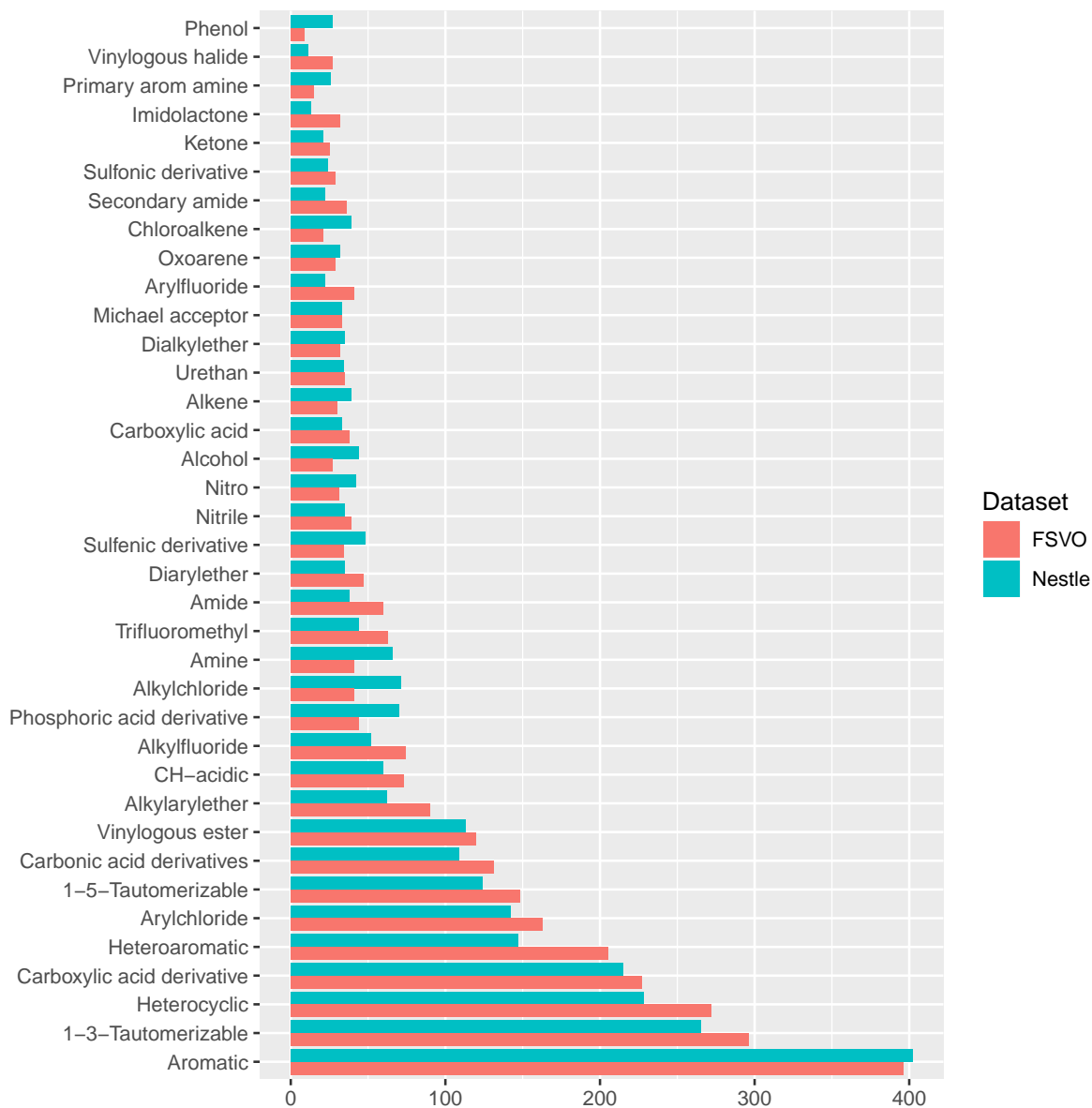


Figure 2: Frequency of functional groups.

233 **Experimental variability versus prediction uncertainty**

234 Duplicated LOAEL values can be found in both databases and there is a substantial number
235 of 155 compounds with more than one LOAEL. These chemicals allow us to estimate
236 the variability of experimental results within individual databases and between databases.
237 Data with *identical* values (at five significant digits) in both databases were excluded from
238 variability analysis, because it is likely that they originate from the same experiments.

239 **Intra database variability**

240 Both databases contain substances with multiple measurements, which allow the determi-
241 nation of experimental variabilities. For this purpose we have calculated the mean LOAEL
242 standard deviation of compounds with multiple measurements. Mean standard deviations
243 and thus experimental variabilities are similar for both databases.

244 The Nestlé database has 567 LOAEL values for 445 unique structures, 93 compounds have
245 multiple measurements with a mean standard deviation (-log₁₀ transformed values) of 0.32
246 (0.56 mg/kg_bw/day, 0.56 mmol/kg_bw/day) (P. Mazzatorta et al. (2008), Figure 3).

247 The FSVO database has 493 rat LOAEL values for 381 unique structures, 91 compounds
248 have multiple measurements with a mean standard deviation (-log₁₀ transformed values) of
249 0.29 (0.57 mg/kg_bw/day, 0.59 mmol/kg_bw/day) (Figure 3).

250 Standard deviations of both databases do not show a statistically significant difference with
251 a p-value (t-test) of 0.21. The combined test set has a mean standard deviation (-log₁₀
252 transformed values) of 0.33 (0.56 mg/kg_bw/day, 0.55 mmol/kg_bw/day) (Figure 3).

253 **Inter database variability**

254 In order to compare the correlation of LOAEL values in both databases and to establish

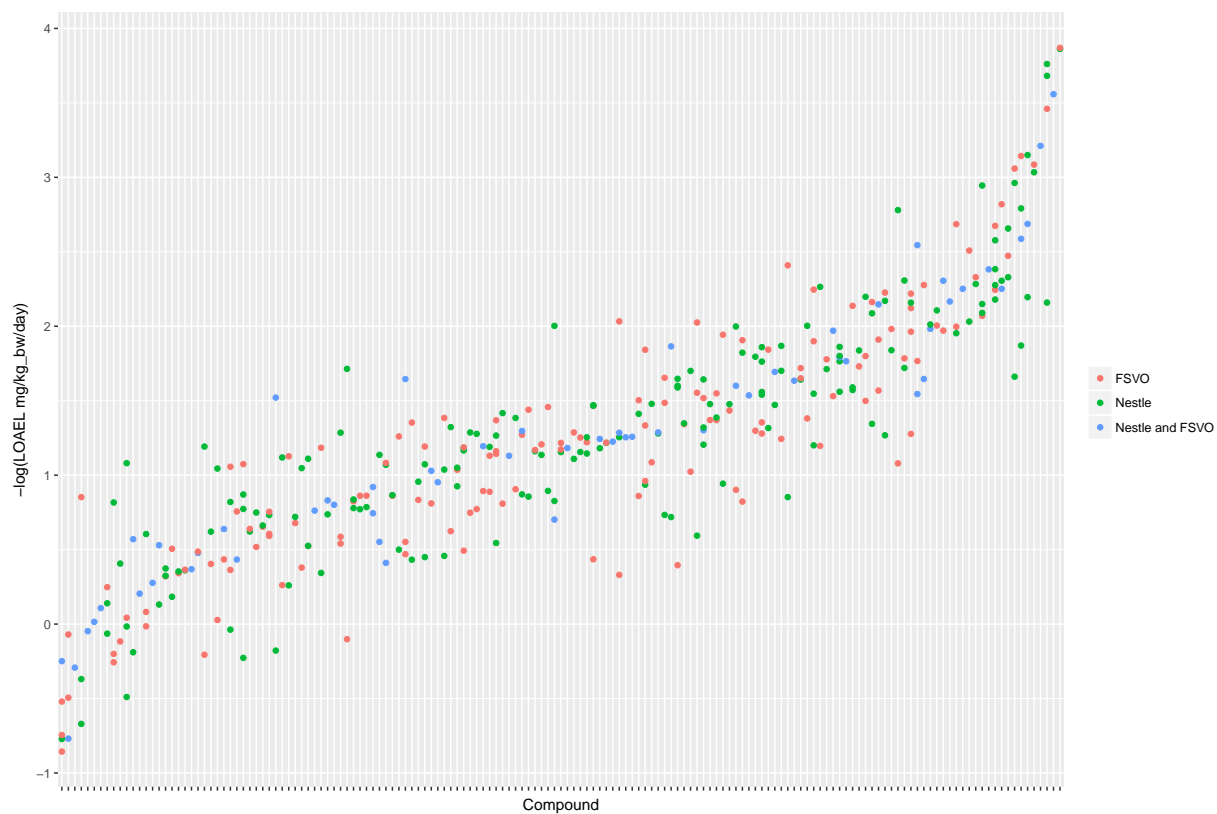


Figure 3: LOAEL distribution and variability of compounds with multiple measurements in both databases. Compounds were sorted according to LOAEL values. Each vertical line represents a compound, and each dot an individual LOAEL value. Experimental variability can be inferred from dots (LOAELs) on the same line (compound).

255 a reference for predicted values, we have investigated compounds, that occur in both
256 databases.

257 Figure 4 depicts the correlation between LOAEL values from both databases. As both
258 databases contain duplicates medians were used for the correlation plot and statistics. It
259 should be kept in mind that the aggregation of duplicated measurements into a single
260 median value hides a substantial portion of the experimental variability. Correlation analysis
261 shows a significant (p-value < 2.2e-16) correlation between the experimental data in both
262 databases with r^2 : 0.52, RMSE: 0.59

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

266 **Local QSAR models**

267 In order to compare the performance of *in silico* read across models with experimental
268 variability we used compounds with multiple measurements as a test set (375 measurements,
269 155 compounds). *lazar* read across predictions were obtained for 155 compounds, 37
270 predictions failed, because no similar compounds were found in the training data (i.e. they
271 were not covered by the applicability domain of the training data).

272 In 100% of the test examples experimental LOAEL values were located within the 95%
273 prediction intervals.

274 Figure 5 shows a comparison of predicted with experimental values. Most predicted values
275 were located within the experimental variability.

276 Correlation analysis was performed between individual predictions and the median of
277 experimental data. All correlations are statistically highly significant with a p-value <

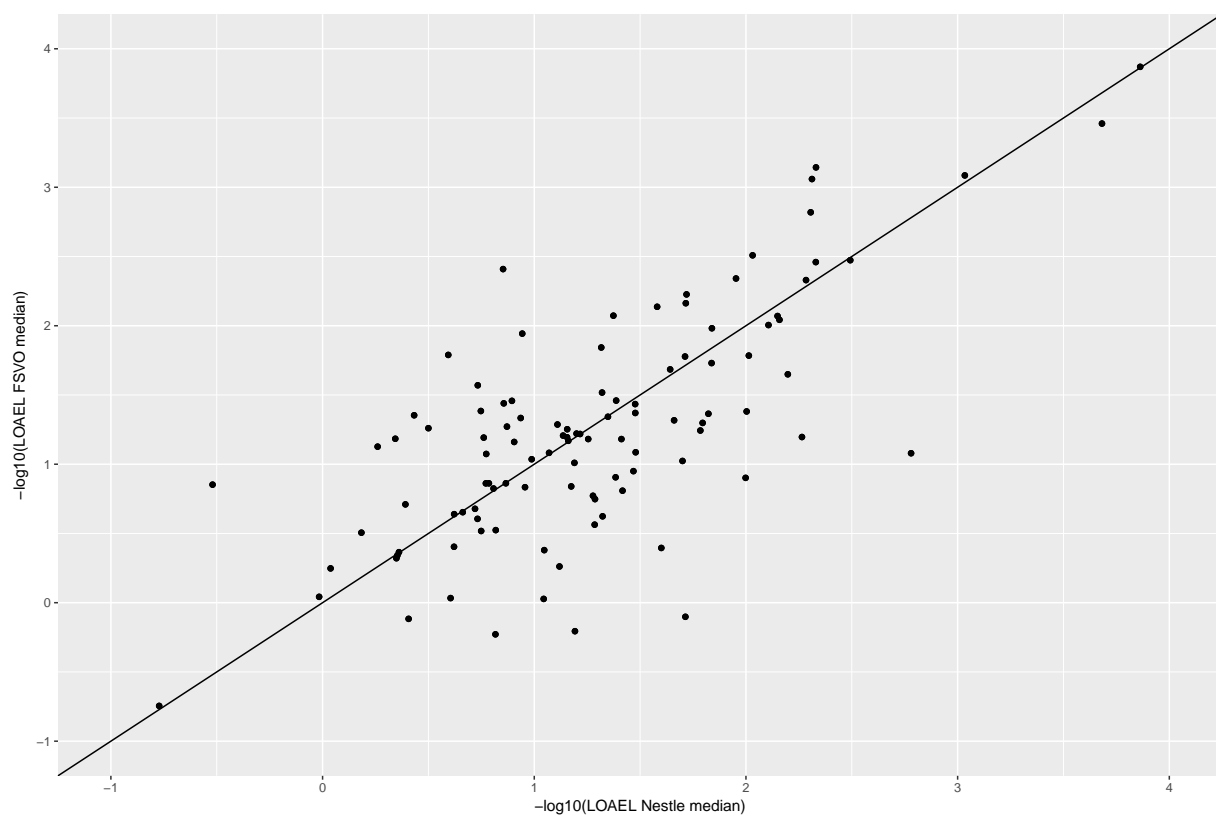


Figure 4: Correlation of median LOAEL values from Nestlé and FSVO databases. Data with identical values in both databases was removed from analysis.

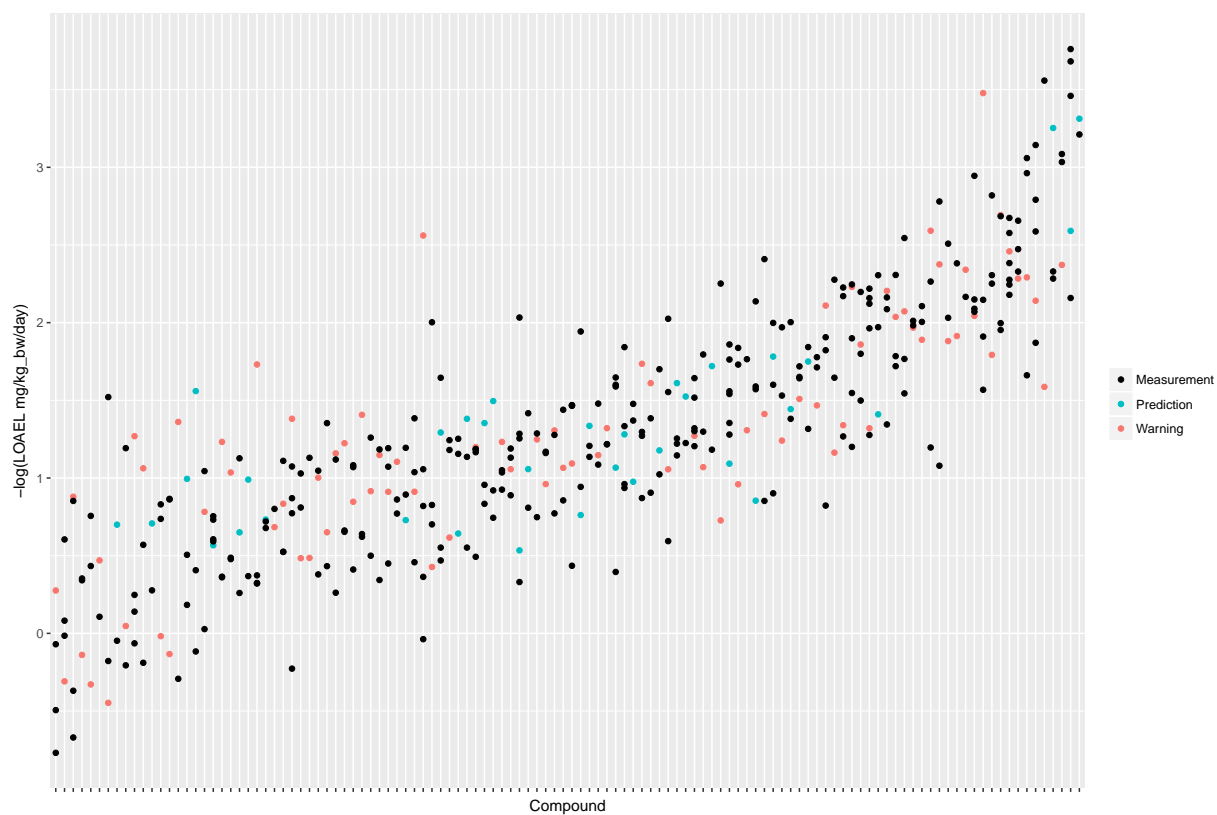


Figure 5: Comparison of experimental with predicted LOAEL values. Each vertical line represents a compound, dots are individual measurements (blue), predictions (green) or predictions far from the applicability domain, i.e. with warnings (red).

278 2.2e-16. These results are presented in Figure 6 and Table 2. Please bear in mind that
279 the aggregation of multiple measurements into a single median value hides experimental
280 variability.

Table 1: Comparison of model predictions with experimental variability.

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

281 For a further assessment of model performance three independent 10-fold cross-validations
282 were performed. Results are summarised in Table 2 and Figure 7. All correlations of
283 predicted with experimental values are statistically highly significant with a p-value <
284 2.2e-16. This was observed for compounds close and more distant to the applicability
285 domain.

Table 2: Results (mean and standard deviation) from 50 independent 10-fold crossvalidations

Predictions	r^2	RMSE	Nr. predicted
AD close	0.6 ± 0.04	0.58 ± 0.02	97 ± 4
AD distant	0.43 ± 0.01	0.8 ± 0.01	380 ± 5
All	0.46 ± 0.01	0.76 ± 0.01	477 ± 4

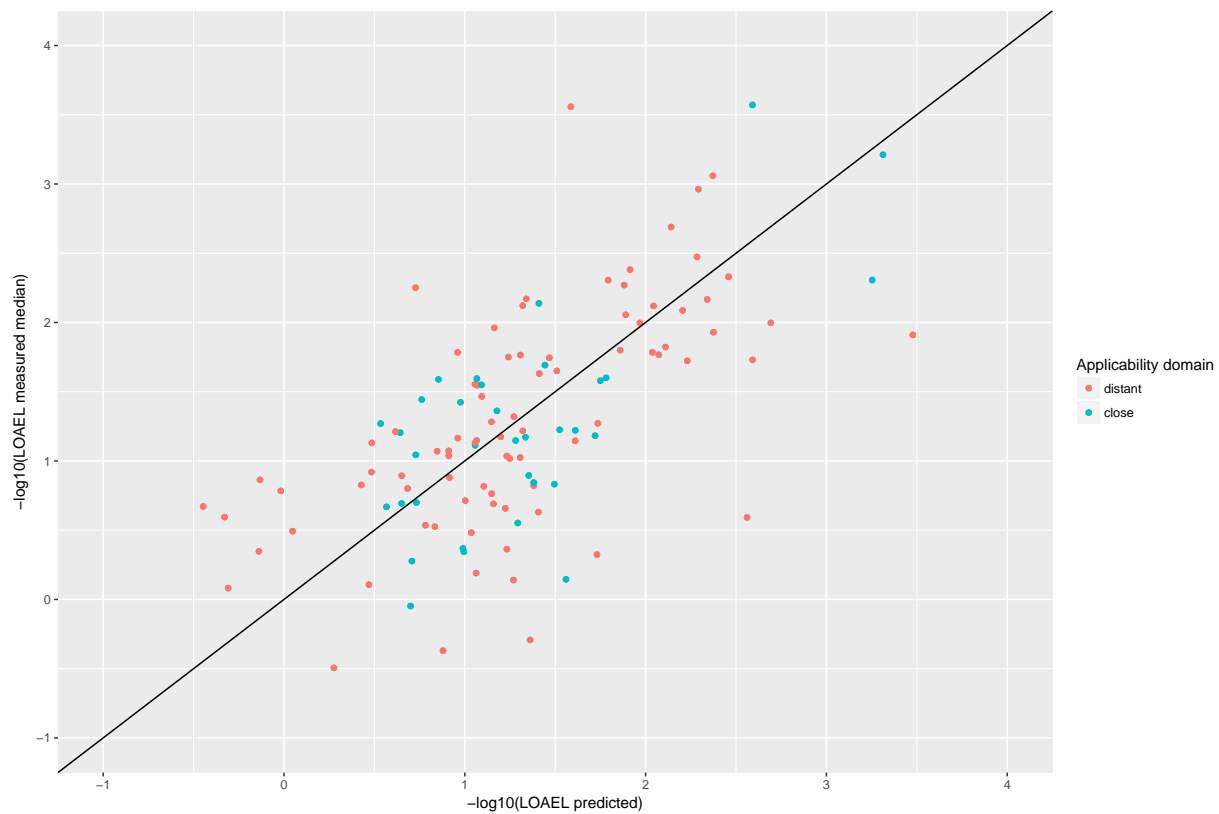


Figure 6: Correlation of experimental with predicted LOAEL values (test set). Green dots indicate predictions close to the applicability domain (i.e. without warnings), red dots indicate predictions far from the applicability domain (i.e. with warnings).

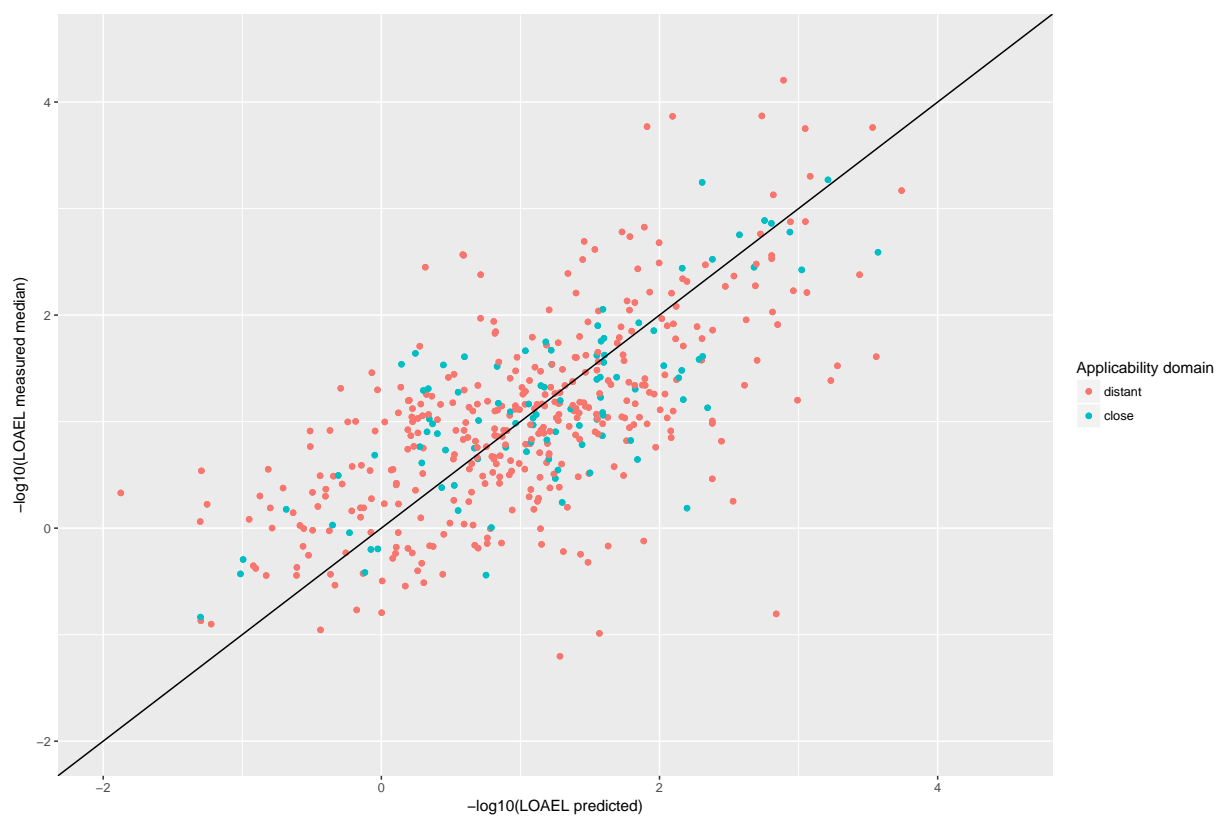


Figure 7: Correlation of predicted vs. measured values from a randomly selected crossvalidation with MP2D fingerprint descriptors and local random forest models.

286 Discussion

287 It is currently acknowledged that there is a strong need for toxicological information on
288 the multiple thousands of chemicals to which human may be exposed through food. These
289 include for example many chemicals in commerce, which could potentially find their way
290 into food (Stanton and Krusezewski 2016, Fowler, Savage, and Mendez (2011)), but also
291 substances migrating from food contact materials (Grob et al. 2006), chemicals generated
292 over food processing (Cotterill et al. 2008), environmental contaminants as well as inherent
293 plant toxicants (Schilter, Constable, and Perrin 2013). For the vast majority of these
294 chemicals, no toxicological data is available and consequently insight on their potential
295 health risks is very difficult to obtain. It is recognized that testing all of them in standard
296 animal studies is neither feasible from a resource perspective nor desirable because of ethical
297 issues associated with animal experimentation. In addition, for many of these chemicals,
298 risk may be very low and therefore testing may actually be irrelevant. In this context,
299 the identification of chemicals of most concern on which limited resource available should
300 focused is essential and computational toxicology is thought to play an important role for
301 that.

302 In order to establish the level of safety concern of food chemicals toxicologically not
303 characterized, a methodology mimicking the process of chemical risk assessment, and
304 supported by computational toxicology, was proposed (Schilter et al. 2014). It is based on
305 the calculation of margins of exposure (MoE) that is the ratio between the predicted chronic
306 toxicity value (LOAEL) and exposure estimate. The level of safety concern of a chemical
307 is then determined by the size of the MoE and its suitability to cover the uncertainties
308 of the assessment. To be applicable, such an approach requires quantitative predictions
309 of toxicological endpoints relevant for risk assessment. The present work focuses on the
310 prediction of chronic toxicity, a major and often pivotal endpoint of toxicological databases

311 used for hazard identification and characterization of food chemicals.

312 In a previous study, automated read-across like models for predicting carcinogenic potency
313 were developed. In these models, substances in the training dataset similar to the query
314 compounds are automatically identified and used to derive a quantitative TD50 value.
315 The errors observed in these models were within the published estimation of experimental
316 variability (Lo Piparo et al. 2014). In the present study, a similar approach was applied
317 to build models generating quantitative predictions of long-term toxicity. Two databases
318 compiling chronic oral rat lowest adverse effect levels (LOAEL) as reference value were
319 available from different sources. Our investigations clearly indicated that the Nestlé and
320 FSVO databases are very similar in terms of chemical structures and properties as well
321 as distribution of experimental LOAEL values. The only significant difference that we
322 observed was that the Nestlé one has larger amount of small molecules, than the FSVO
323 database. For this reason we pooled both databases into a single training dataset for read
324 across predictions.

325 An early review of the databases revealed that 155 out of the 671 chemicals available in
326 the training datasets had at least two independent studies/LOAELs. These studies were
327 exploited to generate information on the reproducibility of chronic animal studies and
328 were used to evaluate prediction performance of the models in the context of experimental
329 variability. Considerable variability in the experimental data was observed. Study design
330 differences, including dose selection, dose spacing and route of administration are likely
331 explanation of experimental variability. High experimental variability has an impact on
332 model building and on model validation. First it influences model quality by introducing
333 noise into the training data, secondly it influences accuracy estimates because predictions
334 have to be compared against noisy data where “true” experimental values are unknown. This
335 will become obvious in the next section, where comparison of predictions with experimental

336 data is discussed. The data obtained in the present study indicate that **lazar** generates
337 reliable predictions for compounds within the applicability domain of the training data
338 (i.e. predictions without warnings, which indicates a sufficient number of neighbors with
339 similarity > 0.5 to create local random forest models). Correlation analysis shows that
340 errors (RMSE) and explained variance (r^2) are comparable to experimental variability of
341 the training data.

342 Predictions with a warning (neighbor similarity < 0.5 and > 0.2 or weighted average predic-
343 tions) are more uncertain. However, they still show a strong correlation with experimental
344 data, but the errors are $\sim 20\text{-}40\%$ larger than for compounds within the applicability domain
345 (Figure 6 and Table 2). Expected errors are displayed as 95% prediction intervals, which
346 covers 100% of the experimental data. The main advantage of lowering the similarity
347 threshold is that it allows to predict a much larger number of substances than with more
348 rigorous applicability domain criteria. As each of this prediction could be problematic, they
349 are flagged with a warning to alert risk assessors that further inspection is required. This
350 can be done in the graphical interface (<https://lazar.in-silico.ch>) which provides intuitive
351 means of inspecting the rationales and data used for read across predictions.

352 Finally there is a substantial number of chemicals (37), where no predictions can be made,
353 because no similar compounds in the training data are available. These compounds clearly
354 fall beyond the applicability domain of the training dataset and in such cases predictions
355 should not be used. In order to expand the domain of applicability, the possibility to design
356 models based on shorter, less than chronic studies should be studied. It is likely that more
357 substances reflecting a wider chemical domain may be available. To predict such shorter
358 duration endpoints would also be valuable for chronic toxicity since evidence suggest that
359 exposure duration has little impact on the levels of NOAELs/LOAELs (Zarn, Engeli, and
360 Schlatter 2011, Zarn, Engeli, and Schlatter (2013)).

361 **lazar** predictions

362 Table 1, Table 2, Figure 5, Figure 6 and Figure 7 clearly indicate that **lazar** generates
363 reliable predictions for compounds within the applicability domain of the training data
364 (i.e. predictions without warnings, which indicates a sufficient number of neighbors with
365 similarity > 0.5 to create local random forest models). Correlation analysis (Table 1, Table 2)
366 shows, that errors ($RMSE$) and explained variance (r^2) are comparable to experimental
367 variability of the training data.

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

378 **Summary**

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

384 our suggested workflow includes always the visual inspection of the chemical structures of
385 the neighbors selected by the model. Indeed it will strength the prediction confidence (if
386 the input structure looks very similar to the neighbors selected to build the model) or it can
387 drive to the conclusion to use read-across with the most similar compound of the database
388 (in case not enough similar compounds to build the model are present in the database).

389 **References**

390 Bender, Andreas, Hamse Y. Mussa, Robert C. Glen, and Stephan Reiling. 2004. “Molecular
391 Similarity Searching Using Atom Environments, Information-Based Feature Selection, and
392 a Naïve Bayesian Classifier.” *Journal of Chemical Information and Computer Sciences* 44
393 (1): 170–78. doi:10.1021/ci034207y.

394 Cotterill, J.V., M.Q. Chaudry, W. Matthews, and R. W. Watkins. 2008. “In Silico Assessment
395 of Toxicity of Heat-Generated Food Contaminants.” *Food Chemical Toxicology*, no. 46(6):
396 1905–18.

397 ECHA. 2008. “Guidance on Information Requirements and Chemical Safety Assessment,
398 Chapter R.6: QSARs and Grouping of Chemicals.” ECHA.

399 EFSA. 2014. “Rapporteur Member State Assessment Reports Submitted for the EU Peer
400 Review of Active Substances Used in Plant Protection Products.” [http://dar.efsa.europa.
401 eu/dar-web/provision](http://dar.efsa.europa.eu/dar-web/provision).

402 EFSA. 2016. “Guidance on the Establishment of the Residue Definition for Dietary
403 Assessment: EFSA Panel on Plant Protect Products and Their Residues (PPR).” *EFSA
404 Journal*, no. 14: 1–12.

405 Fowler, B., S. Savage, and B. Mendez. 2011. “White Paper: Protecting Public Health in

406 the 21st Century: The Case for Computational Toxicology.” ICF International, Inc.icfi.com.

407 Grob, K., M. Biedermann, E. Scherbaum, M. Roth, and K. Rieger. 2006. “Food Contam-
408 ination with Organic Materials in Perspective: Packaging Materials as the Largest and
409 Least Controlled Source? A View Focusing on the European Situation.” *Crit. Rev. Food.*
410 *Sci. Nutr.*, no. 46: 529–35. doi:10.1080/10408390500295490.

411 Gütlein, Martin, Andreas Karwath, and Stefan Kramer. 2012. “CheS-Mapper - Chem-
412 ical Space Mapping and Visualization in 3D.” *Journal of Cheminformatics* 4 (1): 7.
413 doi:10.1186/1758-2946-4-7.

414 Health Canada. 2016. [https://www.canada.ca/en/health-canada/services/chemical-substances/
415 chemicals-management-plan.html](https://www.canada.ca/en/health-canada/services/chemical-substances/chemicals-management-plan.html).

416 Kuhn, Max. 2008. “Building Predictive Models in R Using the Caret Package.” *J. of Stat.*
417 *Soft.*

418 Lo Piparo, E., A. Maunz, C. Helma, D. Vorgrimmler, and B. Schilter. 2014. “Automated and
419 Reproducible Read-Across Like Models for Predicting Carcinogenic Potency.” *Regulatory*
420 *Toxicology and Pharmacology*, no. 70: 370–78.

421 Lo Piparo, E., A. Worth, A. Manibusan, C. Yang, B. Schilter, P. Mazzatorta, M.N. Jacobs,
422 H. Steinkelner, and L. Mohimont. 2011. “Use of Computational Tools in the Field of Food
423 Safety.” *Regulatory Toxicology and Pharmacology*, no. 60(3): 354–62.

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

427 Mazzatorta, Paolo, Manuel Dominguez Estevez, Myriam Coulet, and Benoit Schilter. 2008.
428 “Modeling Oral Rat Chronic Toxicity.” *Journal of Chemical Information and Modeling* 48

429 (10): 1949–54. doi:[10.1021/ci8001974](https://doi.org/10.1021/ci8001974).

430 OBoyle, Noel M, Michael Banck, Craig A James, Chris Morley, Tim Vandermeersch, and
431 Geoffrey R Hutchison. 2011. “Open Babel: An Open Chemical Toolbox.” *Journal of*
432 *Cheminformatics* 3 (1). Springer Science and Business Media: 33. doi:[10.1186/1758-2946-3-](https://doi.org/10.1186/1758-2946-3-33)
433 [33](https://doi.org/10.1186/1758-2946-3-33).

434 OECD. 2015. “Fundamental and Guiding Principles for (Q)SAR Analysis of Chemicals
435 Carcinogens with Mechanistic Considerations Monograph 229 ENV/JM/MONO(2015)46.”
436 In *Series on Testing and Assessment No 229*.

437 Schilter, B., R. Benigni, A. Boobis, A. Chiodini, A. Cockburn, M.T. Cronin, E. Lo Piparo,
438 S. Modi, Thiel A., and A. Worth. 2014. “Establishing the Level of Safety Concern for
439 Chemicals in Food Without the Need for Toxicity Testing.” *Regulatory Toxicology and*
440 *Pharmacology*, no. 68: 275–98.

441 Schilter, B., A. Constable, and I. Perrin. 2013. “Naturally Occurring Toxicants of Plant
442 Origin: Risk Assessment and Management Considerations.” In *Food Safety Management:*
443 *A Practical Guide for Industry*, edited by Y. Motarjemi, 45–57. Elsevier.

444 Stanton, K., and F.H. Krusezewski. 2016. “Quantifying the Benefits of Using Read-Across
445 and in Silico Techniques to Fullfill Hazard Data Requirements for Chemical Categories.”
446 *Regulatory Toxicology and Pharmacology*, no. 81: 250–59. doi:[10.1016/j-yrtph.2016.09.004](https://doi.org/10.1016/j.yrtph.2016.09.004).

447 US EPA. 2011. “Fact Sheets on New Active Ingredients.”

448 Weininger, David. 1988. “SMILES, a Chemical Language and Information System. 1.
449 Introduction to Methodology and Encoding Rules.” *Journal of Chemical Information and*
450 *Computer Sciences* 28 (1): 31–36. doi:[10.1021/ci00057a005](https://doi.org/10.1021/ci00057a005).

451 WHO. 2011. “Joint FAO/WHO Meeting on Pesticide Residues (JMPR) Publications.”

452 <http://www.who.int/foodsafety/publications/jmpr-monographs/en/>.

453 Zarn, J.A., B.E. Engeli, and J.R. Schlatter. 2011. “Study Parameters Influencing NOAEL
454 and LOAEL in Toxicity Feeding Studies for Pesticides: Exposure Duration Versus Dose
455 Decrement, Dose Spacing, Group Size and Chemical Class.” *Regul. Toxicol. Pharmacol.*,
456 no. 61: 243–50.

457 ———. 2013. “Characterization of the Dose Decrement in Regulatory Rat Pesticide Toxicity
458 Feeding Studies.” *Regul. Toxicol. Pharmacol.*, no. 67: 215–20.