nano-lazar: Read across predictions for nanoparticle toxicities with calculated and measured properties

Christoph Helma[1]*, Micha Rautenberg[1], Denis Gebele[1]

[1] in silico toxicology gmbh, Basel, Switzerland

Correspondence*:
Christoph Helma, in silico toxicology gmbh, Rastatterstr. 41, CH-4057 Basel, Switzerland
helma@in-silico.ch

2 ABSTRACT

The lazar framework for read across predictions was expanded for the prediction of nanoparticle toxicities, and a new methodology for calculating nanoparticle descriptors from core and coating structures was implemented. In order to compare nanoparticle descriptor sets and local regression algorithms 60 independent crossvalidation experiments were performed for the Protein Corona dataset obtained from the eNanoMapper database. The best RMSE and rˆ2 results were obtained with protein corona descriptors and the weighted random forest algorithm, but its 95% prediction interval is significantly less accurate than for models using simpler descriptor sets (measured and calculated nanoparticle properties). The most accurate prediction intervals were obtained with measured nanoparticle properties with RMSE and rˆ2 values that show no statistical significant difference ($p < 0.05$) to the protein corona descriptors. Calculated descriptors are interesting for cheap and fast high-throughput screening purposes, random forest models have significantly lower rˆ2 values, but RMSE and prediction intervals are comparable to protein corona and nanoparticle random forest models.

16 Keywords: nanoparticle, toxicity, QSAR, read-across, predictive toxicology

1 INTRODUCTION

Read across is a commonly used approach for the risk assessment of chemicals. Read across procedures are based on the assumption that similar compounds cause similar biological effects. In order to estimate the activity of a novel compound a researcher will search for similar compounds with known biological activities and deduce the activity of the new compound from this data. In order to make the read across procedure reproducible, traceable and objective the authors of this paper have developed a computer program (`lazar`, (Maunz et al. 2013)) that automates the risk assessment process. The objective of the current study was to extend `lazar` for the risk assessment of nanomaterials.

The concept of chemical *similarity* is the key idea behind all read across procedures. But similarity is not an intrinsic property of substances, it can be defined in different ways and the utility and performance of similarity measures depends on each specific use case.

Structural similarity is most frequently used in the risk assessment of compounds with a well defined chemical structure. Structural similarity definitions are obviously not directly applicable to nanomaterials,

29 because they lack a well defined structure. It is however relatively straightforward to adapt other concepts,
30 e.g. similarity in terms of chemical properties or in terms of biological effects. Compared to structural
31 similarity, which can be calculated directly from chemical structures, these similarity definitions depend
32 on actual measurements, which makes their estimation more expensive and time consuming. For this
33 reason we have developed a novel concept of structural similarity for nanomaterials, which is based on the
34 chemical fingerprints of core and coating materials. According to our knowledge, this is the first time that
35 nanoparticle toxicities have been predicted successfully from calculated properties alone.

36 In order to estimate the utility of various similarity concepts for nanomaterials, we have performed model
37 building and validation experiments for models based on

38 • *structural similarity* (using on core and coating fingerprints)
39 • *property similarity* (using on measured nanoparticle properties)
40 • *biological similarity* (using serum protein interaction data)

41 and the local regression algorithms

42 • weighted average
43 • weighted partial least squares
44 • weighted random forests

45 In addition we intend to address the important topic of *reproducible research* with this publication. It is in
46 our experience frequently impossible to reproduce computational experiments for a variety of reasons, e.g.

47 • publications lack important details about algorithms
48 • publications do not provide access to the data that has been used
49 • authors use proprietary software that does not disclose its algorithms with all necessary details
50 • original software, libraries and operating systems are outdated and not available anymore

51 Our attempt to address these problems is to provide a self contained environment that contains all software
52 and data for the experiments presented in this manuscript. It contains also a build system for the manuscript,
53 that pulls results and figures directly from validation experiments (similar to the `R knitr` package (Xie
54 2015)).

55 A self-contained system with the compiled manuscript and all libraries and external programs required
56 for repeating the validation experiments is publicly available as a `docker` image from DockerHub
57 (`https://hub.docker.com/r/insilicotox/nano-lazar-paper`). Apart from repeating
58 the experiments for this paper this image can also be used for extending the system, testing other descriptor
59 and modelling algorithms and comparing validation results with the current benchmark.

60 Source code for the manuscript and validation experiments has been published under a GPL3 license at
61 Github (`https://github.com/opentox/nano-lazar-paper`). The `lazar` framework library
62 has been published under the same license (`https://github.com/opentox/lazar`).

63 A graphical webinterface for `nano-lazar` model predictions and validation results is publicly
64 accessible at `https://nano-lazar.in-silico.ch`, source code for the GUI can be obtained
65 from `https://github.com/enanomapper/nano-lazar`.

66 Github and DockerHub repositories are tagged with `nano-lazar-paper` to identify the software
67 version that corresponds to the published paper. As this project is under continuous development, it is
68 likely that some of the algorithms will change in the future. In this case it is relatively straightforward to

69 identify differences with the versioning system or to use the submitted version as benchmark for further
70 developments.

2 METHODS

71 The following sections give a high level overview about `nano-lazar` algorithms. Readers interested in
72 unambiguous algorithm definitions can refer to the source code links in the text.

2.1 Datasets

74 Nanoparticle characterisations and toxicities were mirrored from the eNanoMapper database
75 (Jeliazkova et al. 2015) via its REST API (`https://github.com/opentox/lazar/blob/`
76 `nano-lazar-paper.review/lib/import.rb#L9-L118`). At present only the *Net cell*
77 *association* endpoint of the *Protein corona* dataset, has a sufficient number of examples (121) to create and
78 validate read-across models, all other public nanoparticle endpoints have less than 20 examples, which
79 makes them unsuitable for local QSAR modelling and crossvalidation experiments.

2.2 Algorithms

81 For this study we have adapted the modular `lazar` (*la*zy structure *a*ctivity *r*elationships) read across
82 framework (Maunz et al. 2013) for nanoparticle model development and validation.

83 lazar was originally developed for small molecules with a defined chemical structure and uses chemical
84 fingerprints for the identification of similar compounds (*neighbors*). Nanoparticles in contrast do not
85 have clearly defined chemical structures, but they can be characterised by their composition (core and
86 coatings), measured properties (e.g. size, shape, physicochemical properties) or the interaction with
87 biological macromolecules. Within `nano-lazar` we use these properties for the identification of similar
88 nanoparticles (*neighbors*) and as descriptors for local QSAR models.

89 `nano-lazar` makes read-across predictions with the following basic workflow: For a given nanoparticle
90 `lazar`

91 • searches in the database for similar nanoparticles (*neighbors*) with experimental toxicity data,
92 • builds a local QSAR model with these neighbors and
93 • uses this model to predict the activity of the query compound.

94 This procedure resembles an automated version of *read across* predictions in toxicology, in machine
95 learning terms it would be classified as a *k-nearest-neighbor* algorithm (`https://github.com/`
96 `opentox/lazar/blob/nano-lazar-paper.review/lib/model.rb#L180-L257`).

97 Apart from this basic workflow `nano-lazar` is completely modular and allows the researcher to use
98 arbitrary algorithms for similarity searches and local QSAR modelling. Within this study we are using and
99 comparing the following algorithms:

2.2.1 Nanoparticle descriptors

101 In order to find similar nanoparticles and to create local QSAR models it is necessary to characterize
102 nanoparticles by descriptors. In this study we are using three types of descriptors:

Structural descriptors Calculated molecular fingerprints for core and coating compounds (MOLPRINT 2D fingerprints (Bender et al. 2004), *MP2D*, `https://github.com/opentox/lazar/blob/nano-lazar-paper.review/lib/nanoparticle.rb#L17-L21`)

Physico-chemical nanoparticle properties Measured nanoparticle properties from the eNanoMapper database (*P-CHEM*)

Biological nanoparticle properties Protein interaction data from the eNanoMapper database (*Proteomics*)

Nanoparticle fingerprints are a novel development for the characterisation of nanoparticles with well defined core and coating compounds. In this case it is possible to create molecular fingerprints for all of these compounds and use the union of these fingerprints as nanoparticle fingerprint. Based on our experience with small molecules we have selected MOLPRINT 2D fingerprints (Bender et al. 2004), which typically outperform predefined fingerprints (e.g. $MACCS$, $FP4$) for QSAR purposes. Despite its simplicity the concept works surprisingly well (see validation results) and enables toxicity predictions without measured properties. This can be useful e.g. for fast and cheap nanoparticle toxicity screening programs.

2.2.2 Feature selection

Calculated MP2D fingerprints are used without feature selection, as preliminary experiments have shown, that feature selection deteriorates the overall performance of read-across models (which is in agreement with our observations on small molecules).

Nanoparticle properties in the eNanoMapper database have not been measured for the purpose of read across and QSAR modelling. For this reason the database contains a lot of features that are irrelevant for toxicity. In preliminary experiments we have observed that using all available features for similarity calculations leads to neighbor sets that are unsuitable for local QSAR models, because large numbers of irrelevant features override the impact of features that are indeed relevant for toxicity.

For this reason we use the `lazar` concept of *activity specific similarities* (Maunz et al. 2013), by selecting only those features that correlate with a particular toxicity endpoint (Pearson correlation p-value < 0.05). This reduced set of *relevant features* is used for similarity calculations and local QSAR models (`https://github.com/opentox/lazar/blob/nano-lazar-paper.review/lib/feature_selection.rb#L6-L26`). Apart from being computationally cheaper, simple filter methods pose also a lower risk of overfitting than more aggressive feature selection methods (e.g. forward selection, backwards elimination). As local models are built with the `R caret` package which uses feature selection internally there is no requirement for extremely small descriptor sets at this stage.

For crossvalidation experiments feature selection is repeated separately for each crossvalidation fold, to avoid overfitted models (Gütlein et al. 2013).

2.2.3 Neighbor identification

For binary features (MP2D fingerprints) we are using the union of core and coating fingerprints to calculate the Tanimoto/Jaccard index and a similarity threshold of $sim > 0.1$ (`https://github.com/opentox/lazar/blob/nano-lazar-paper.review/lib/similarity.rb#L18-L20`).

For quantitative features (P-CHEM, Proteomics) we use the reduced set of relevant features to calculate the *weighted cosine similarity* of their scaled and centered relevant feature vectors, where the contribution of each feature is weighted by its Pearson correlation coefficient with the toxicity

144 endpoint. A similarity threshold of $sim > 0.5$ was used for the identification of neighbors for
145 local QSAR models (`https://github.com/opentox/lazar/blob/nano-lazar-paper.`
146 `review/lib/similarity.rb#L37-L49`).

147 In both cases nanoparticles that are identical to the query particle are eliminated from neighbors
148 to obtain unbiased predictions in the presence of duplicates. (`https://github.com/opentox/`
149 `lazar/blob/nano-lazar-paper.review/lib/model.rb#L180-L257`).

2.2.4 Local QSAR models and predictions

151 For read-across predictions local QSAR models for a query nanoparticle are build from the set of similar
152 nanoparticles (*neighbors*).

153 In this investigation we are comparing three local regression algorithms:

154 • weighted local average (*WA*, `https://github.com/opentox/lazar/blob/nano-lazar-paper.`
155   `review/lib/regression.rb#L6-L16`)
156 • weighted partial least squares regression (*PLS*, `https://github.com/opentox/lazar/`
157   `blob/nano-lazar-paper.review/lib/caret.rb#L7-L78`)
158 • weighted random forests (*RF*, `https://github.com/opentox/lazar/blob/nano-lazar-paper`
159   `review/lib/caret.rb#L7-L78`)

160 In all cases neighbor contributions are weighted by their similarity to the query particle. The weighted
161 local average algorithm serves as a simple and fast benchmark algorithm, whereas partial least squares and
162 random forests are known to work well for a variety of QSAR problems. Partial least squares and random
163 forest models use the `caret` R package (Kuhn 2008). Models are trained with the default `caret` settings,
164 optimizing the number of PLS components or number of variables available for splitting at each RF tree
165 node by bootstrap resampling.

166 Finally the local model is applied to predict the activity of the query nanoparticle. The RMSE of
167 bootstrapped model predictions is used to construct 95% prediction intervals at 1.96*RMSE (`https://`
168 `github.com/opentox/lazar/blob/nano-lazar-paper.review/lib/caret.rb#L55-L77`).
169 Prediction intervals are not available for the weighted average algorithm, as it does not use internal
170 validation.

171 If PLS/RF modelling or prediction fails, the program resorts to using the weighted average method.

2.2.5 Applicability domain

173 The applicability domain of `lazar` models is determined by the diversity of the training data. If no
174 similar compounds are found in the training data (either because there are no similar nanoparticles or
175 because similarities cannot be determined due to the lack of measured properties) no predictions will
176 be generated. Warnings are also issued, if local QSAR model building or model predictions fail and the
177 program has to resort to the weighted average algorithm (`https://github.com/opentox/lazar/`
178 `blob/nano-lazar-paper.review/lib/model.rb#L180-L257`).

179 Each prediction is accompanied with a list of neighbors and their similarities, which are clearly displayed
180 in the graphical user interface for the inspection by a toxicological expert. Apart from indicating the
181 applicability domain, the neighbor list clearly shows the rationale for the prediction, and allows the expert
182 to reject predictions e.g. when neighbors act via different mechanisms.

183 The accuracy of local model predictions is indicated by the 95% prediction interval, which
184 is derived from the internal `caret` validation (`https://github.com/opentox/lazar/`
185 `blob/nano-lazar-paper.review/lib/caret.rb#L55-L77`). Query substances close to the
186 applicability domain (many neighbors with high similarity) will have a narrower prediction interval than
187 substances with a larger distance (few neighbors with low similarity).

2.2.6 Validation

189 For validation purposes we use results from 5 repeated 10-fold crossvalidations with independent
190 training/test set splits for each descriptor/algorithm combination (`https://github.com/opentox/`
191 `lazar/blob/nano-lazar-paper.review/lib/crossvalidation.rb#L85-L93`). Feature
192 selection is performed for each validation fold separately to avoid overfitting. For the same reason we
193 do not use a fixed random seed for training/test set splits. This leads to slightly different results for each
194 repeated crossvalidation run, but it allows to estimate the variability of validation results due to random
195 training/test splits.

196 In order to identify significant differences between validation results, outcomes (RMSE, r^2,
197 correct 95% prediction interval) are compared by ANOVA analysis, followed by Tukey multiple
198 comparisons of means (`https://github.com/enanomapper/nano-lazar-paper/blob/`
199 `nano-lazar-paper.review/scripts/cv-statistics.rb`).

200 Please note that recreating validations (e.g. in the Docker image) will not lead to exactly the same results,
201 because crossvalidation folds are created randomly to avoid overfitting for fixed training/test set splits.

202 These five 10-fold crossvalidations are assigned to the final model, which is build from the complete
203 training data. This validated model is used for further predictions, e.g. from the graphical webinterface.

2.3 Availability

205 **Public webinterface** `https://nano-lazar.in-silico.ch`
206 **`lazar` framework** `https://github.com/opentox/lazar` (source code)
207 **`nano-lazar` GUI** `https://github.com/enanomapper/nano-lazar` (source code)
208 **Manuscript** `https://github.com/opentox/nano-lazar-paper` (source code for the
209 manuscript and validation experiments)
210 **Docker image** `https://hub.docker.com/r/insilicotox/nano-lazar-paper/` (container
211 with manuscript, validation experiments, `lazar` libraries and third party dependencies)

3 RESULTS

212 The *Protein corona dataset* contains 121 Gold and Silver particles that are characterized by physchem
213 properties (*P-CHEM*) and their interaction with proteins in human serum (*Proteomics*). In addition *MP2D*
214 fingerprints were calculated for core and coating compounds with defined chemical structures.

215 Five repeated crossvalidations with independent training/test set splits were performed for the descriptor
216 classes

217 • *MP2D* fingerprints (calculated, binary)
218 • *P-CHEM* properties (measured, quantitative)
219 • *Proteomics* data (measured, quantitative)
220 • *P-CHEM* and *Proteomics* data combined (measured, quantitative)

221 and the local regression algorithms

222 • local weighted average (*WA*)
223 • local weighted partial least squares regression (*PLS*)
224 • local weighted random forests (*RF*)

225 Results of these experiments are summarized in Table 1. Figure 1, Figure 2 and Figure 3
226 show the correlation of predictions with measurements for *MP2D*, *P-CHEM* and *Proteomics*
227 random forests models. Correlation plots for all descriptors and algorithms are available as
228 supplementary material (`https://github.com/enanomapper/nano-lazar-paper/tree/`
229 `nano-lazar-paper.review/figures`). Table 2 lists *P-CHEM* properties of the Protein Corona
230 dataset and their correlation with the *Net Cell Association* endpoint.

4 DISCUSSION

231 Table 1 summarizes the results from five independent crossvalidations for all descriptor/algorithm
232 combinations. The best results in terms of $RMSE$ and R^2 were obtained with *Proteomics* descriptors
233 and local weighted *random forest* models. There are however six models without statistically significant
234 differences in terms of $RMSE$ and five models in terms of r^2. The most accurate 95% prediction intervals
235 were obtained with *P-CHEM* descriptors and *random forest* models, this models does not differ significantly
236 from the best $RMSE$ and r^2 results.

4.1 Descriptors

238 In terms of descriptors the best overall results were obtained with *Proteomics* descriptors. This is in
239 agreement with previous findings from other groups (Walkey et al. 2014, Liu et al. (2015), Papa et al.
240 (2016)). It is however interesting to note that the prediction intervals are significantly more inaccurate than
241 those from other descriptors and the percentage of measurements within the prediction interval is usually
242 lower than 90% instead of the expected 95%.

243 Using *P-CHEM* descriptors in addition to *Proteomics* does not lead to improved models, instead we
244 observe an increased sensitivity towards training/test set splits (crossvalidation variability) and *random
245 forest* results perform even significantly poorer than *Proteomics* descriptors alone.

246 *P-CHEM* descriptors alone perform surprisingly well, especially in combination with local *random forest*
247 models, which does not show statistically significant differences to the best *Proteomics* model. On average
248 more than 95% of the measurements fall within the 95% prediction interval, with significantly better results
249 than for *Proteomics* descriptors. A summary of *P-CHEM* descriptors can be found in Table 2.

250 All *MP2D* models have poorer performance in terms of r^2, but the *random forest* model does not differ
251 significantly in terms of $RMSE$ and measurements within the prediction interval.

4.2 Algorithms

253 With the exception of *P-CHEM/Proteomics* descriptors *random forests* models perform
254 better than *partial least squares* and *weighted average* models with significant differences
255 for *MP2D* and *P-CHEM* descriptors (detailed pairwise comparisons are available in the
256 supplementary material `https://github.com/enanomapper/nano-lazar-paper/blob/`
257 `nano-lazar-paper.review/results/`). Interestingly the simple *weighted average* algorithm

258 shows no significant difference to the best performing model for the *Proteomics* and *P-CHEM/Proteomics*
259 descriptors.

4.3 Interpretation and practical applicability

261   Although *random forest* models with *Proteomics* descriptors have the best performance in terms of
262 $RMSE$ and r^2, the accuracy of the 95% prediction interval is significantly lower than for *MP2D* and
263 *P-CHEM* models (detailed pairwise comparisons in the supplementary material).

264   These problems seem to originate from internal `caret` optimisation and validation algorithms which
265 underestimate $RMSE$ values, that are used to calculate the prediction interval (see Algorithm section).
266 The observation that the *weighted average* algorithm, which does not use `caret`, performs comparatively
267 well for *Proteomics* descriptors, supports this interpretation.

268   Our initial suspicion was that an unfavourable ratio between descriptors (785 before feature selection,
269 129 after feature selection) and training examples (121) causes this problem. $Random forest$ and
270 $partial least squares$ algorithms are on the other hand robust against a large number of descriptors and
271 `caret` returns very realistic $RMSE$ values for *MP2D* fingerprints with a similar number of independent
272 variables (100). For this reason it is presently still unclear, why prediction intervals for *Proteomics*
273 descriptors are more inaccurate than for other descriptor types.

274   *P-CHEM random forest* models have the most accurate prediction interval and the $RMSE$ and r^2
275 performance is comparable to the $Proteomics$ model, although they utilize a much lower number of
276 descriptors (20 before feature selection, 10 after feature selection). The main advantage from a practical
277 point of view is that predictions of novel nanoparticles require a much lower amount of measurements
278 than with $Proteomics$ data (although this argument may become obsolete with new high throughput
279 techniques).

280   *MP2D* fingerprint descriptors are interesting from a practical point of view, because they do not require
281 any measurements of nanoparticle properties. They need however defined chemical structures for core and
282 coating compounds, which makes this approach infeasible for nanoparticle classes like carbon nanotubes.
283 The resulting models do not differ significantly from the best results in terms of prediction accuracy
284 ($RMSE$, measurements within prediction interval), but are significantly lower in terms of explained model
285 variance (r^2). For practical purposes one may argue that the primary objective of read across models is to
286 make accurate predictions (low $RMSE$, accurate prediction interval) and not to explain the model variance
287 (r^2). For this reason we consider r^2 performance as secondary compared to $RMSE$ and prediction interval
288 accuracies.

289   Currently a couple of QSAR studies with global models have been published for the same dataset Walkey
290 et al. (2014), Liu et al. (2015), Papa et al. (2016)], but unfortunately their results are not directly comparable,
291 because we report results for the complete dataset with 121 Gold and Silver particles, while other authors
292 report results for a subset of Gold particles.

293   (Walkey et al. 2014) report leave-one-out (LOO) and 4-fold crossvalidation ($4CV$) results for 105 Gold
294 particles. They obtained the best results (LOO r^2 0.86, 4CV r^2 0.63) with partial least squares models,
295 protein corona data with four additional physicochemical parameters and jackknife parameter selection.
296 Parameter selection was performed by crossvalidation, but it is unclear if parameters were selected on
297 the complete dataset prior to LOO/4CV or separately for each LOO/4CV model. Performance wise the
298 findings are roughly in agreement with our results. Assuming that feature selection was performed within
299 crossvalidation folds we would expect 10-fold crossvalidation results between LOO and $4CV$ results.

300 According to the authors the model developed for Gold compounds have little predictivity for Silver
301 compounds, but a separate Silver model gave LOO r^2 of 0.79. $RMSE$ values are not available, although
302 they are in our opinion more relevant for the predictive toxicology use case than r^2 values (prediction error
303 vs explained model variance).

304   (Liu et al. 2015) report a 4CV r^2 of 0.843 for 84 Gold compounds using ϵ-support vector machines
305 (ϵ-SVM) with 6 serum proteins and zeta potential as descriptors. Descriptors were selected with sequential
306 forward floating selection (*SFFS*). The methodological descriptions do not indicate explicitly, if feature
307 selection was performed on the complete dataset or within 4CV folds. Judging from Figure 2 of this
308 paper and the Methods section we have the strong impression that feature selection was performed prior
309 to crossvalidation, which increases the likelihood of overfitted models, especially for aggressive feature
310 selection schemes like *SFFS*. The 4CV $r2$ of 0.843 is clearly higher than our results, but it remains unclear,
311 if the superior performance is due to better algorithms, a smaller more "regression friendly" dataset or
312 overfitted models. Again we would have preferred $RMSE$ values for comparison purposes, which are
313 unfortunately not available.

314   (Papa et al. 2016) developed global models for 84 Gold compounds with eleven algorithms and reported
315 r^2 and $RMSE$ values for training set retrofitting, leave-one-out crossvalidation (*LOO*) and stratified
316 external test set predictions (64 particles training set, 20 particles test set). There was little difference
317 between good performing models (PPR, EARTH, SVM-linear, SVM-radial, MLR, PLS) and the authors
318 conclude that Projection Pursuit Regression (PPR) gives the most robust models (LOO r^2 0.81, $RMSE$
319 1.01, external r^2 0.79, $RMSE$ 1.01).

320   TODO feature selection The best results have been obtained with TODO get full paper and discuss

5   CONCLUSION

321 We have performed 60 independent crossvalidation experiments for the Protein Corona dataset obtained
322 from the eNanoMapper database in order to identify the best combination of descriptors for nanoparticle
323 read across predictions. The best RMSE and r^2 results were obtained with protein corona descriptors and
324 the weighted random forest algorithm, but the 95% prediction interval is significantly less accurate than
325 that of models with simpler descriptor sets (measured and calculated nanoparticle properties). The most
326 accurate prediction intervals were obtained with measured nanoparticle properties with RMSE and r^2
327 values that show no statistical significant difference (p < 0.05) to the protein corona descriptors. Calculated
328 descriptors are interesting for cheap and fast high-throughput screening purposes, they have significantly
329 lower r^2 values than the best results, but RMSE and prediction intervals show no significant difference to
330 the best results of our investigation.

331   For practical purposes we suggest to use nanoparticle properties when measurements are available and the
332 newly developed nanoparticle fingerprints for screening purposes without physicochemical measurements.
333 Both models have been implemented with a graphical user interface which is publicly available at `https:`
334 `//nano-lazar.in-silico.ch`.

6   CONFLICT OF INTEREST STATEMENT

335 The authors declare that the research was conducted in the absence of any commercial or financial
336 relationships that could be construed as a potential conflict of interest.

7 AUTHOR CONTRIBUTIONS

337 CH was responsible for the design and implementation of the `nano-lazar` libraries, the validation studies
338 and the text of this manuscript. DG and MR participated as scientific programmers in the development of
339 `nano-lazar` libraries and in the validation experiments. They are the authors of the `nano-lazar` GUI
340 and REST interfaces and contributed to the manuscript with critical revisions and proofreading.

8 FUNDING

9 TABLES

Table 1 *P-CHEM* properties of the *Protein corona* dataset. Features correlating with the *Net cell association* endpoint (*relevant features*) are indicated by bold letters.

Property	Medium	Unit
Localized Surface Plasmon Resonance (LSPR) index	-	
Localized Surface Plasmon Resonance (LSPR) index	**Human serum**	
LSPR peak position (nm)	-	nm
Polydispersity index	**-**	nm
Polydispersity index	Human serum	nm
Core size	**-**	nm
Autot (ICP-AES)	**Human serum**	$nmol$
Total surface area (SAtot)	Human serum	cm^2
Protein density	Human serum	ug/cm^2
Total protein (BCA assay)	**Human serum**	ug
ZETA POTENTIAL	**-**	mV
ZETA POTENTIAL	**Human serum**	mV
Z-Average Hydrodynamic Diameter	-	nm
Z-Average Hydrodynamic Diameter	**Human serum**	nm
Volume Mean Hydrodynamic Diameter	-	nm
Volume Mean Hydrodynamic Diameter	**Human serum**	nm
Number Mean Hydrodynamic Diameter	-	nm
Number Mean Hydrodynamic Diameter	Human serum	nm
Intensity Mean Hydrodynamic Diameter	-	nm
Intensity Mean Hydrodynamic Diameter	**Human serum**	nm

Table 2   Results from five independent crossvalidations for various descriptor/algorithm combinations. Best results (mean of 5 crossvalidations) are indicated by bold letters, statistically significant ($p < 0.05$) different results by italics. Results in normal fonts do not differ significantly from best results.

Descriptors	Algorithm	RMSE	r^2	% measurements within prediction interval
MP2D	WA	*2.04 2.0 2.02 2.07 2.07*	*0.24 0.27 0.25 0.22 0.22*	NA
MP2D	PLS	*2.14 2.11 2.21 1.99 1.9*	*0.27 0.26 0.26 0.32 0.36*	94 97 91 91 97
MP2D	RF	1.84 1.67 1.68 1.69 1.71	*0.4 0.5 0.49 0.48 0.47*	94 96 96 94 94
P-CHEM	WA	*1.91 1.93 1.91 2.03 2.02*	*0.48 0.47 0.49 0.41 0.42*	NA
P-CHEM	PLS	*2.2 2.33 2.11 2.27 2.21*	*0.34 0.28 0.38 0.31 0.33*	97 92 96 93 91
P-CHEM	RF	1.8 1.82 1.77 1.68 1.86	0.54 0.53 0.56 0.6 0.51	**94 96 97 97 93**
Proteomics	WA	1.94 1.63 1.7 1.61 1.76	0.49 0.64 0.6 0.64 0.57	NA
Proteomics	PLS	1.67 1.63 1.86 1.74 1.8	0.62 0.64 0.53 0.59 0.56	*90 88 84 89 88*
Proteomics	RF	**1.66 1.69 1.81 1.68 1.6**	**0.62 0.61 0.57 0.6 0.65**	*89 89 89 87 89*
P-CHEM Proteomics	WA	1.61 1.56 1.71 1.66 2.41	0.64 0.66 0.6 0.62 0.33	NA
P-CHEM Proteomics	PLS	1.74 1.67 1.78 1.71 2.18	0.6 0.62 0.59 0.61 0.43	*91 90 86 85 86*
P-CHEM Proteomics	RF	*1.78 1.62 1.56 1.82 1.77*	*0.57 0.64 0.66 0.55 0.61*	*88 87 87 89 90*

REFERENCES

345 Bender, Andreas, Hamse Y. Mussa, and Robert C. Glen, and Stephan Reiling. 2004. "Molecular
346 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, Christoph Helma, Andreas Karwath, and Stefan Kramer. 2013. "A Large-Scale Empirical Evaluation of Cross-Validation and External Test Set Validation in (Q)SAR." *Molecular Informatics* 32 (5-6). WILEY-VCH Verlag: 516–28. doi:10.1002/minf.201200134.

Jeliazkova, Nina, Charalampos Chomenidis, Philip Doganis, Bengt Fadeel, Roland Grafström, Barry Hardy, Janna Hastings, et al. 2015. "The ENanoMapper Database for Nanomaterial Safety Information." *Beilstein J. Nanotechnol.*, no. 6: 1609–34. doi:doi:10.3762/bjnano.6.165.

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

Liu, Rong, Wen Jiang, Carl D. Walkey, Warren C. W. Chan, and Yoram Cohen. 2015. "Prediction of Nanoparticles-Cell Association Based on Corona Proteins and Physicochemical Properties." *Nanoscale* 7 (21). The Royal Society of Chemistry: 9664–75. doi:10.1039/C5NR01537E.

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.

Papa, E., J. P. Doucet, A. Sangion, and A. Doucet-Panaye. 2016. "Investigation of the Influence of Protein Corona Composition on Gold Nanoparticle Bioactivity Using Machine Learning Approaches." *SAR and QSAR in Environmental Research* 27 (7): 521–38. doi:10.1080/1062936X.2016.1197310.

Walkey, Carl D., Jonathan B. Olsen, Fayi Song, Rong Liu, Hongbo Guo, D. Wesley H. Olsen, Yoram Cohen, Andrew Emili, and Warren C. W. Chan. 2014. "Protein Corona Fingerprinting Predicts the Cellular Interaction of Gold and Silver Nanoparticles." *ACS Nano* 8 (3): 2439–55. doi:10.1021/nn406018q.

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. http://yihui.name/knitr/.
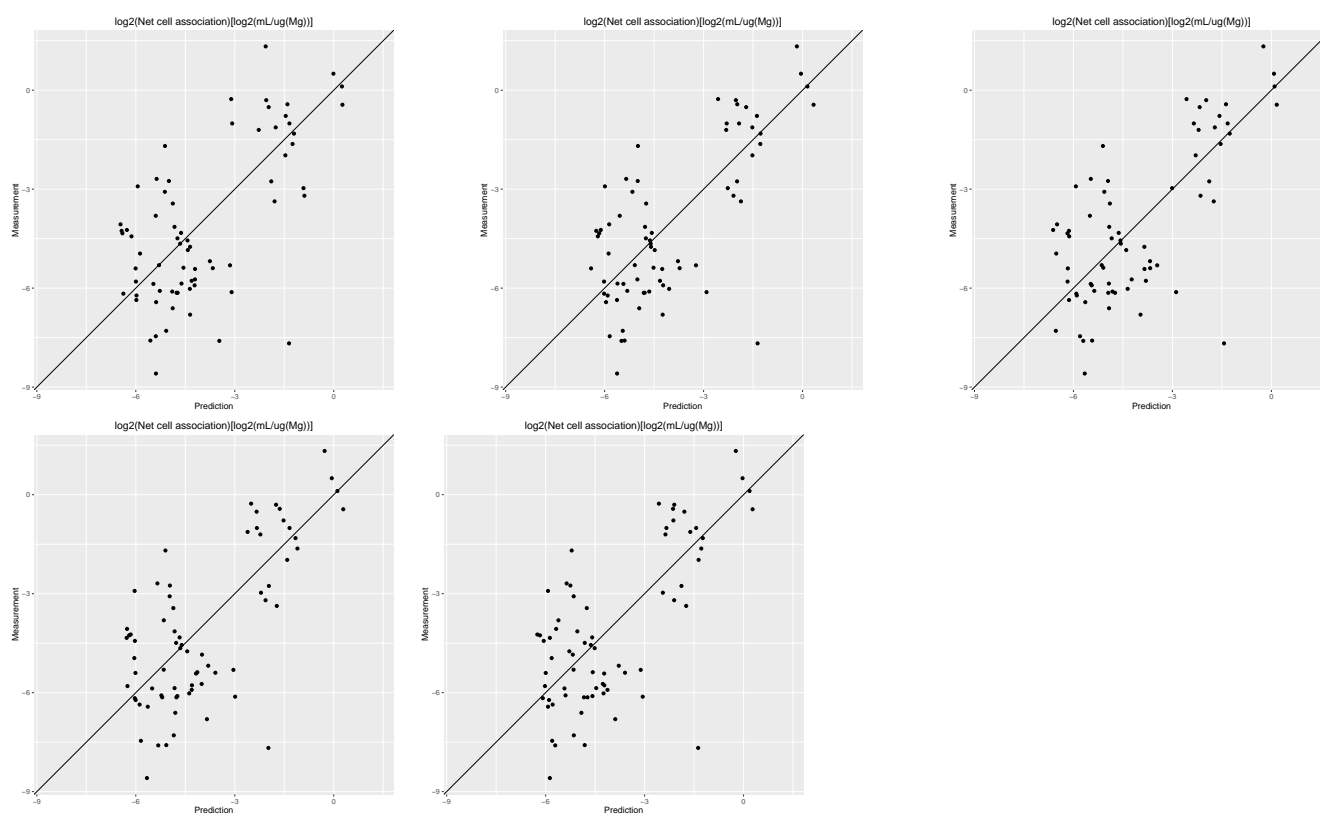
10 FIGURES

Figure 1. Correlation of predicted vs. measured values for five independent crossvalidations with *MP2D* fingerprint descriptors and local *random forest* models
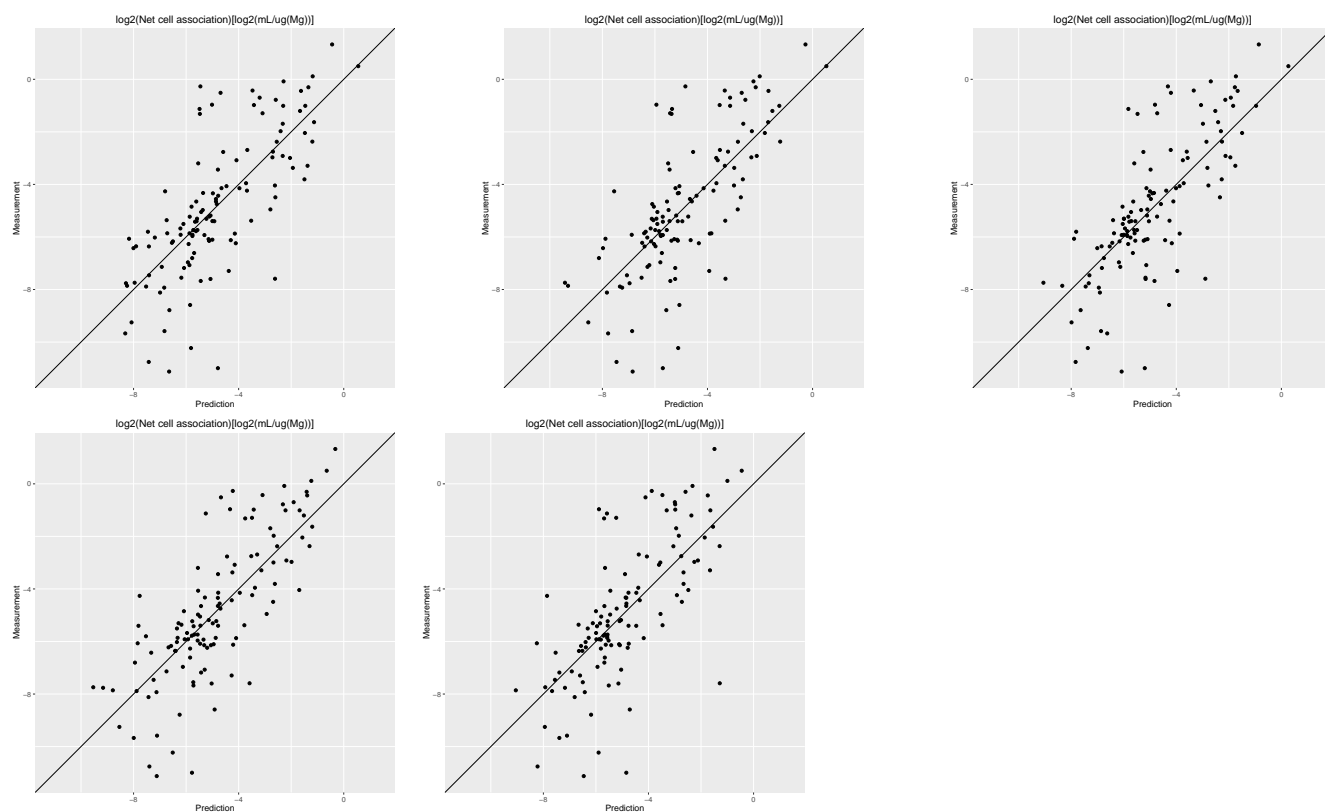
Figure 2. Correlation of predicted vs. measured values for five independent crossvalidations with *P-CHEM* descriptors and local *random forest* models
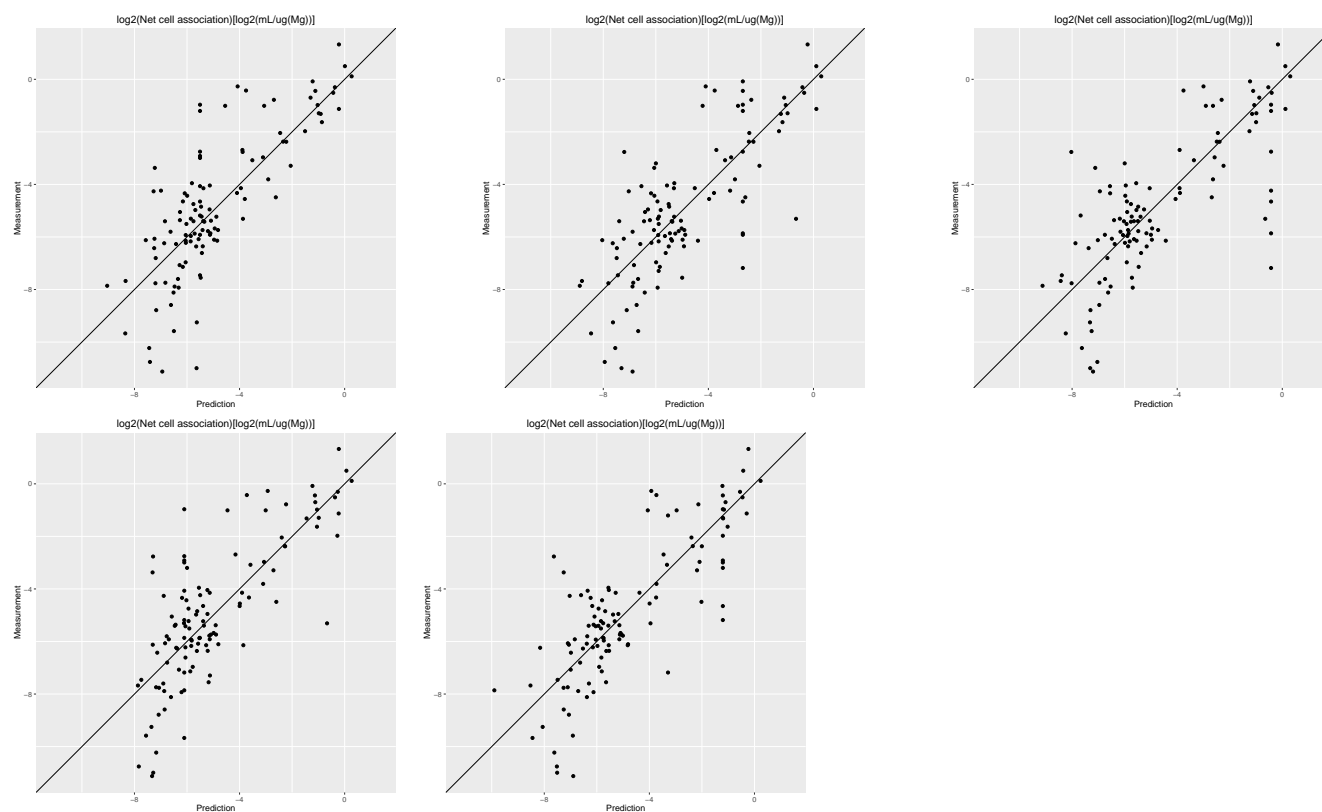
Figure 3. Correlation of predicted vs. measured values for five independent crossvalidations with *Proteomics* descriptors and local *random forest* models