nano-lazar: Validation of read across predictions for nanoparticle toxicities

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

December 12, 2016

[1] in silico toxicology gmbh, Basel, Switzerland

Introduction

Data requirements

Calculation of similarities intersection of physchem descriptors

Experimental data for similar compounds

Use cases

- no nanoparticle information: core+coating properties
- physchem measurements
- proteomics

Objectives

- Evaluate currently available nanoparticle data for read across predictions
- Compare read across predictions based on
- calculated core and coating properties
- measured nanoparticle properties
- nanoparticle protein corona

Reproducible research

With this investigation we intend to provide an example of reproducible research.

This manuscript has been generated by a build system that pulls results and figures directly from validation data. Source code for the manuscript and the associated libraries are publicly available under a GPL license from the GitHub repositories https://github.com/opentox/nano-lazar-paper (manuscript) and https://github.com/opentox/lazar (lazar framework).

For the reproduction of results with exactly the same libraries, dependencies and programs at the time of the manuscript creation we provide additionally a self-contained docker image at DockerHub https://hub.docker.com/r/insilicotox/nano-lazar-paper/.

Please note that recreating validations will not lead to exactly the same results, because we deliberately avoid setting a predefined random seed for crossvalidation folds, in order to avoid overfitting for fixed training/test set splits.

TODO: GUI @ https://nano-lazar.in-silico.ch

Methods

Datasets

Nanoparticle characterisations and toxicities were mirrored from the eNanoMapper database (Jeliazkova et al. 2015) via its REST API.

Algorithms

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

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

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

- searches in a database for similar nanoparticles (*neighbors*) with experimental toxicity data,
- builds a local QSAR model with these neighbors and
- uses this model to predict the 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 nano-lazar is completely modular and allows the researcher to use arbitrary algorithms for similarity searches and local QSAR modelling. Within this study we are using and comparing the following algorithms:

Nanoparticle descriptors

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

- Calculated molecular fingerprints for core and coating compounds (MOLPRINT 2D fingerprints (Bender et al. 2004), *MP2D*)
- Measured nanoparticle properties from the eNanoMapper database (*P-CHEM*)
- Protein interaction data from the eNanoMapper database (*Proteomics*)

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), which leads to a set of *relevant features*. This reduced feature set is used for similarity calculations and local QSAR models. For crossvalidation experiments feature selection is repeated separately for each crossvalidation fold, to avoid overfitted models [@????].

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 $sim > 0.1$.

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 endpoint. A similarity threshold of $sim > 0.5$ is used for the identification of neighbors for local QSAR models.

In both cases nanoparticles that are identical to the query particle are eliminated from neighbors to obtain unbiased predictions in the presence of duplicates.

Local QSAR models and predictions

For read-across predictions local QSAR models for a query nanoparticle are build with similar nanoparticles (*neighbors*).

In this investigation we are comparing three local regression algorithms:

- weighted local average (WA)
- weighted partial least squares regression (PLS)
- weighted random forests (RF)

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

Finally the local model is applied to predict the activity of the query nanoparticle. The RMSE of bootstrapped model predictions is used to construct 95% prediction intervals at 1.96*RMSE. Prediction intervals are not available for the weighted average algorithm, as it does not use internal validation,

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

Applicability domain

The applicability domain of lazar models is determined by the diversity of the training data. If no similar compounds are found in the training data (either because there are no similar nanoparticles or because similarities cannot

be determined du to the lack of mesured properties) no predictions will be generated. Warnings are also issued, if local QSAR model building or model predictions fail and the program has to resort to the weighted average algorithm.

The accuracy of local model predictions is indicated by the 95% prediction interval.

Validation

For validation purposes we use results from 3 repeated 10-fold crossvalidations with independent training/test set splits. Feature selection is performed separately for each training dataset to avoid overfitting. For the same reason we do not use a fixed random seed for training/test set splits. This leads to slightly different results for each repeated crossvalidation run, but it allows to estimate the variability of validation results due to random training/test splits.

In order to identify significant differences between validation results, outcomes (RMSE, r^2, correct 95% prediction interval) are compared by ANOVA analysis, followed by Tukey multiple comparisons of means.

Results

The first step step was to determine the toxicity endpoints currently available in the eNanoMapper database that have sufficient data for the creation and validation of read across models. Table **??** summarizes the endpoints and data points that are currently available in eNanoMapper.

Table 1: Substances per endpoint.

Dataset	Endpoint	Nanoparticles
NanoWiki	Concentration in cell	4
NanoWiki	Log Reciprocal EC50	17
NanoWiki	LDH Release	5
NanoWiki	DNA in Tail	5
NanoWiki	Metabolic Activity	5
NanoWiki	Toxicity Classifier	9
NanoWiki	Percentage Viable Cells	4
NanoWiki	Concentration in culture medium	1
Protein Corona	Net cell association	121
Protein Corona	log2(Net cell association)	121
MARINA	TNF-alpha	6
MARINA	% cell viability	6
MODENA	Cell Viability Assay EC25	1
MODENA	Cell Viability Assay EC50	1

Dataset	Endpoint	Nanoparticles
MODENA	Cell Viability Assay SLOPE EC50	41
MODENA	LDH Release Assay EC25	10
MODENA	LDH Release Assay EC50	10
MODENA	LDH Release Assay SLOPE EC50	11
MODENA	ATP Assay EC25	8
MODENA	ATP Assay EC50	8
MODENA	ATP Assay SLOPE EC50	8
MODENA	MTT Assay EC25	10
MODENA	MTT Assay EC50	10
MODENA	MTT Assay SLOPE EC50	10

In order to obtain meaningful and statistically relevant results from crossvalidation experiments we need at least 100 examples per endpoint. In our experience feature selection and local model building frequently fails for smaller datasets (especially within crossvalidation folds) because too few examples are available and crossvalidation results depend more on training/test set splits than on the performance of individual algorithms. This general observation was confirmed by attempts to validate models for the *Cell Viability* endpoint of the MODENA dataset with 41 examples and 4 independent features. In these cases global models may be preferable over local read-across models, but these models will have a narrow applicability domain.

At present only the *Net cell association* endpoint of the *Protein corona* dataset, has a sufficient number of examples to create and validate read-across models. It contains 121 Gold and Silver particles that are characterized by physchem properties (*P-CHEM*) and their interaction with proteins in human serum (*Proteomics*). In addition *MP2D* fingerprints were calculated for core and coating compounds with defined chemical structures.

Table 2: *P-CHEM* properties of the *Protein corona* dataset

Abbreviation	Description
TODO	Example

Three repeated crossvalidations with independent training/test set splits were performed for the descriptor classes

- *MP2D* fingerprints (calculated, binary)
- *P-CHEM* properties (measured, quantitative)
- *Proteomics* data (measured, quantitative)
- *P-CHEM* and *Proteomics* data combined (measured, quantitative)

and the local regression algorithms

- local weighted average (*WA*)
- local weighted partial least squares regression (*PLS*)
- local weighted random forests (*RF*)

Results of these experiments are summarized in Table **??**. Figures Figure **??** and Figure **??** show the correlation of predictions with measurements for the *P-CHEM/Proteomics* and *MP2D* random forests models. Correlation plots for all descriptors and algorithms are available in the supplementary material.

Table 3: Repeated crossvalidation results.

Descriptors	Algorithm	RMSE	r^2	% within prediction interval
MP2D	WA	*2.11 2.02 2.07*	*0.18 0.25 0.22*	NA
MP2D	PLS	1.88 2.03 1.86	*0.39 0.3 0.39*	**96 96 96**
MP2D	RF	1.71 1.9 1.71	*0.48 0.35 0.47*	96 94 94
P-CHEM	WA	1.97 1.93 1.99	*0.45 0.47 0.44*	NA
P-CHEM	PLS	*2.27 2.09 2.17*	*0.29 0.38 0.34*	94 93 97
P-CHEM	RF	1.8 1.78 1.71	0.54 0.55 0.59	95 95 97
Proteomics	WA	1.85 1.74 1.85	0.53 0.58 0.53	NA
Proteomics	PLS	*1.69 2.55 1.9*	0.61 0.35 0.53	*88 84 89*
Proteomics	RF	**1.54 1.58 1.64**	**0.67 0.66 0.63**	*89 91 91*
P-CHEM Proteomics	WA	1.67 1.66 1.64	0.61 0.63 0.63	NA
P-CHEM Proteomics	PLS	1.83 1.73 1.57	0.57 0.6 0.66	*88 82 89*
P-CHEM Proteomics	RF	1.95 1.79 2.08	*0.51 0.56 0.47*	*90 90 88*

Discussion

p-chem/proteomics rf best performing mp2d/rf most practical

relevant features features used in local models

calculated vs measured practical applicability prediction interval accuracy

variability of results

Liu paper:

descriptor selection not included in cv!! prediction accuracy != r^2 uses bootstrap and strange r^2 which includes training set performance

all papers: no silver particles

georgia:

why only 84 gold particles (neutrals excluded) text could be clearer unterschied 10cv, 10cv-test is this clustering supervised or unsupervised

mixture of regulatory, (nano)tox and machine learning/stat aspects conceptional overview of BIO descriptors before formal definition statistically significant differences of results (?) liu study overfitted!! (discussion) references, figures sometimes incorrect VIP comes from lui? => choosing preselected proteins == overfitting

which contains TODO Gold and Silver particles that are characterized by physchem properties and their interaction with proteins in human serum. For this dataset we have found TODO (NTUA abstract?) reference studies (Walkey et al. 2014, Liu et al. (2015)).

TODO: literature search

https://scholar.google.com/scholar?q=protein+corona+nanoparticles+qsar&btnG=&hl=en&as_sdt=0%2C5&a

TODO: description of parameters

Conclusion

Acknowledgements

This work was performed as part of the EU FP7 project "Nanomaterials safety assessment: Ontology, database(s) for modelling and risk assessment Development of an integrated multi-scale modelling environment for nanomaterials and systems by design" (Theme NMP.2013.1.3-2 NMP.2013.1.4-1, Grant agreement no: 604134).

References

McDermott et al., 2013; Walkey et al., 2014 Yang et al., 2012; Balbin et al., 2013

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.

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.

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.