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

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

December 8, 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 in terms of measured 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 follows the following basic workflow: For a given nanoparticle lazar

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

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

Apart from this basic workflow 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:

Feature selection

Nanoparticle properties in the eNanoMapper database have not been measured with read across and QSAR modelling in mind. For this reason the database contains a lot of features that are irrelevant for toxicity. Using all available features for similarity calculations leads neighbor sets that are unsuitable for local QSAR models, because large numbers of irrelevant features override the impact of relevant features.

TODO: example, results section?

For this reason we return to the lazar concept of *activity specific similarities* (Maunz et al. 2013), by selecting features that correlate with a particular toxicity endpoint (Pearson correlation p-value < 0.05), which leads to a set of *relevant features*. This procedure is repeated separately for each crossvalidation fold, to avoid overfitted models [@????].

Neighbor identification

Similarity calculations are based on the reduced set of relevant features that correlate well with the toxic effect.

The chemical similarity between two nanoparticles is defined as 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.

A similarity threshold of $sim > 0.5$ is used for the identification of neighbors for local QSAR models. Nanoparticles that are identical to the query particle are eliminated from the neighbors to obtain unbiased predictions in the presence of duplicates.

Local QSAR models and predictions

Only similar nanoparticles (*neighbors*) above the threshold are used for local QSAR models. 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.

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.

The variability of local model predictions is reflected in the prediction interval.

Validation

For validation purposes we use the 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.

Results

Data requirements

The first in our experiments 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	41
MODENA	LDH Release Assay	11
MODENA	ATP Assay	8
MODENA	MTT Assay	10

In order to a threshold of at least 100 examples This criterea is currently fulfilled only by the *Net cell association* endpoint of the *Protein corona* dataset, 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

Repeated crossvalidations

This section presents the results of repeated crossvalidation experiments with nanoparticle read across models for the *Net cell association* endpoint (log2 transformed).

We have investigated the following descriptor classes

- Physchem properties TODO size, shape??
- Proteomics data (TODO erklaeren)

- Physchem properties and proteomics data

and the local regression algorithms

- local weighted average
- local weighted partial least squares regression
- local weighted random forests

Table 2: Repeated crossvalidation results.

Descriptors	Algorithm	RMSE	r^2	% within prediction interval
MP2D	WA	2.04 2.02 2.03	*0.24 0.25 0.24*	NA
MP2D	PLS	1.83 1.97 1.97	*0.41 0.32 0.33*	**99 96 97**
MP2D	RF	1.72 1.78 1.67	0.47 0.44 0.49	96 93 96
Proteomics	WA	1.71 1.79 1.7	0.6 0.56 0.6	NA
Proteomics	PLS	1.7 1.68 1.99	0.6 0.61 0.51	*90 89 89*
Proteomics	RF	2.07 1.62 1.6	0.5 0.64 0.65	*87 90 88*
P-CHEM Proteomics	WA	1.86 1.93 1.63	0.55 0.51 0.64	NA
P-CHEM Proteomics	PLS	1.73 1.76 2.64	0.64 0.59 0.31	*92 85 90*
P-CHEM Proteomics	RF	**1.59 1.58 1.6**	**0.66 0.66 0.66**	*85 89 90*

Discussion

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

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

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.