The analogue bias: pernicious effects also in QSAR

Floriane Montanari[1], Santiago D. Villalba, Gerhard F. Ecker[1]

[1]University of Vienna, Dept. of Pharmaceutical Chemistry, Vienna, Austria
Email to: floriane.montanari@univie.ac.at

https://github.com/sdvillal/manysources

References

[1] Rohrer and Baumann, MUV datasets for virtual screening based on Pubchem bioactivity data, J *Chem Inf Model*., 2009.
[2] Montanari and Ecker, BCRP inhibition: from data collection to ligand-based modeling, *Mol Inf.*, 2014.
[3] Giannini *et al*., E-ring-modified 7-oxyiminomethyl camptothecins: Synthesis and preliminary in vitro and in vivo biological evaluation, *Bioorg Med Chem Lett*., 2008.
[4] An *et al*., Cellular phototoxicity evoked through the inhibition of human ABC transporter ABCG2 by cyclin-dependent kinase inhibitors in vitro, *Pharm Res*., 2009.

Background

Most QSAR datasets are artificially enriched in given chemical families because medicinal chemists typically focus on a certain scaffold and explore substitution patterns around it. In virtual screening, this effect is known as "analogue bias" [1] and leads to overoptimistic results. What happens when such datasets are used to build machine learning models? We focus here on a BCRP inhibition dataset collected from different sources [2] as an example of typical multiple-source QSAR dataset. Other datasets were used and showed similar results (data not shown).

Aim of the study

Use interpretable models (logistic regression on unfolded fingerprints) and clever data splitting strategies to understand:
- the influence of chemical scaffolds on individual predictions
- the influence of the training set on individual predictions
- the impact of analogue bias on learning

A data leaking example from Giannini_2008

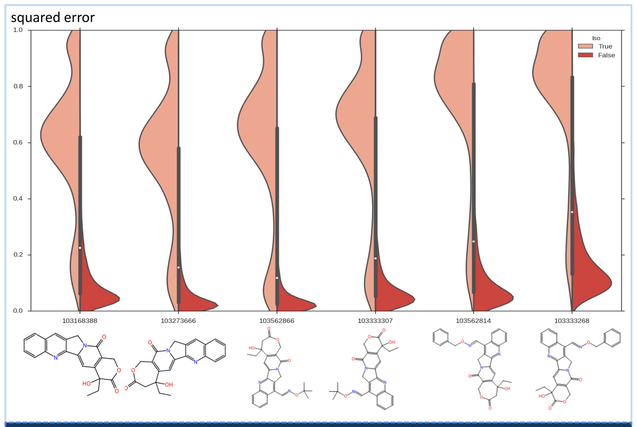


Fig. 3: Distribution of the squared errors for each compound of the source Giannini_2008, evaluated in LSO (light red) or random split (dark red).

All compounds in Giannini_2008 [3] share the same scaffold and are inhibitors of BCRP. The squared error per compound is high in the LSO setting (light red), *i.e*., those compounds are not well predicted by the LSO models. On the contrary, the squared error is small in the RS setting (dark red), *i.e*. the presence of one or two compounds from the source in training allows for easy prediction of the other similar compounds.

Data splitting strategies: random split versus leave-sources-out

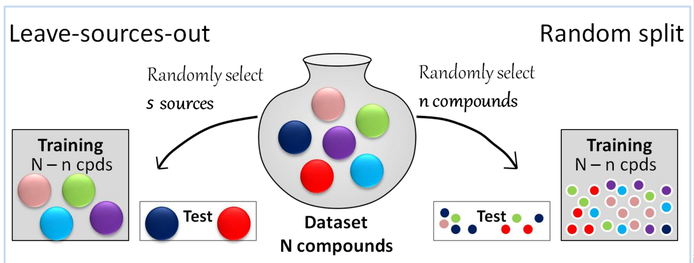


Fig. 1: Splitting scheme to compare random split evaluation versus leave-source-out evaluation. The colors represent each source.

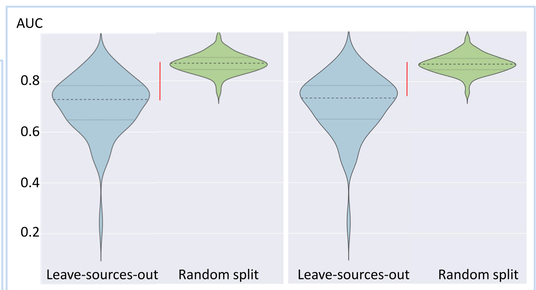


Fig. 2: Distribution of AUC values over 120 splits, Random Forest, CDK descriptors. On the left side, no applicability domain is applied. On the right side, an applicability domain filter is applied.

A source (publication, database entry) corresponds to a natural grouping of QSAR datasets built on open data. We use this grouping to split the data in a "leave-sources-out" (LSO) fashion: random sets of sources are used as external set (Fig. 1). This simulates a prospective, real-world evaluation of the models. We then compare with the traditional random splitting (RS) of compounds.
Results (Fig. 2) show a much lower predictive value in prospective evaluation than in classical evaluation. This result hold for all descriptors sets (Maccs, ECFP8, CDK2D, VolSurf), algorithms (Random Forest, SVM, Logistic regression, naïve Bayes) and multiple-source datasets (BCRP inhibition, P-gp inhibition, hERG inhibition, mutagenicity) used.

An_2008: learning a scaffold

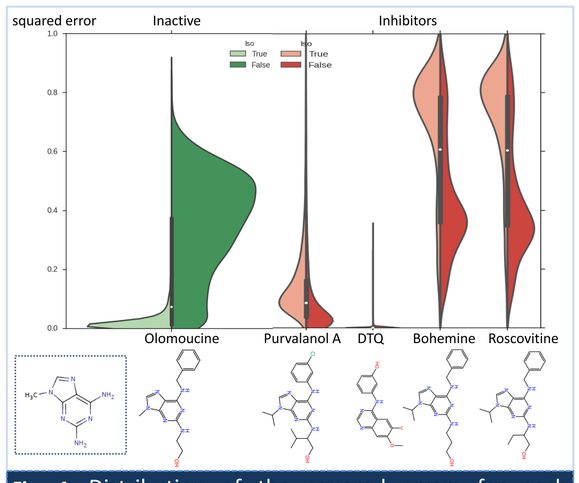


Fig. 4: Distribution of the squared errors for each compound of the source An_2008, evaluated in LSO (light color) or random split (dark color). The common scaffold is given in the box.

	Olomoucine	Purvalanol	Bohemine	Roscovitine
Prediction LSO	☺	☺	☹	☹
Prediction RS	☹	☺	☺	☺
Contains c(CN)(cc)cc	yes	no	yes	yes
Contains n(c(c)N)c(N)n	yes	yes	yes	yes

The weights of two substructures occurring in this source (Fig. 5) explain the squared error distributions (Fig. 4). Further confirmation of the influence of Bohemine and Roscovitine on Olomoucine predictions is shown in Fig. 6.

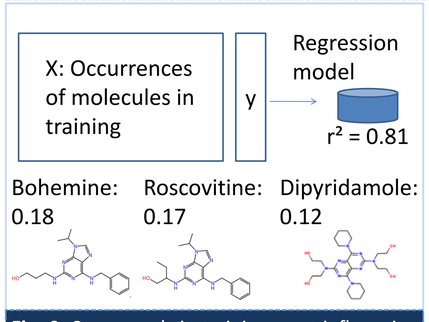


Fig. 6: Compounds in training most influencing the squared error of olomoucine in RS models.

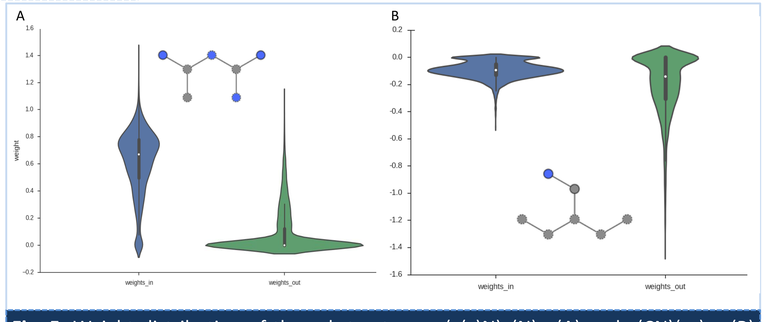
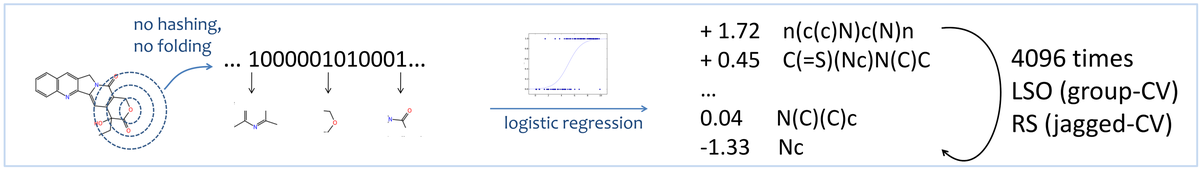


Fig. 5: Weight distribution of the substructures n(c(c)N)c(N)n (A) and c(CN)(cc)cc (B) when the source An_2008 is part of the training set (blue) or not (green).

Interpretable models



Here, we want to dig into the dynamics between models, training set composition and evaluation. Morgan fingerprints are built using the RDKit library. Each substructure is directly encoded without hashing or folding, which avoids collisions. Logistic regression models are trained, allowing to retrieve the weights given to each substructure. Many models are built on both LSO- and RS-like cross-validations: in LSO-CV, each source occurs once in test; in RS-CV, each compound occurs once in test; the sizes of the folds are the same in LSO and RS.

Conclusions

The effect of applying machine learning techniques on datasets suffering from analogue bias is two-fold. First, the **usual evaluation metrics do not reflect anymore realistic estimations for the predictive capabilities of the model**, as the chemical similarity between families of compounds leads to information leaking. This was shown using the leave-sources-out evaluation and by looking at the Giannini_2008 example. Second, **the models themselves overfit the over-represented scaffolds**, leading to mispredictions of activity cliffs. This was shown using the An_2008 example.