Nicolas Bosc, Francis Atkinson, Eloy Felix, Anna Gaulton, Anne Hersey and Andrew R. Leach
nbosc@ebi.ac.uk
ChEMBL group, European Bioinformatics Institute, EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, United Kingdom

EMBL-EBI

How does conformal prediction perform on a wide range of ChEMBL datasets?

Introduction

Conformal prediction (CP) has emerged as a solution to assess confidence from the predictions themselves[1]. Directly derived from QSAR in its implementation, CP does not require much more CPU time and is straightforward to implement. Moreover, the method allows one to pick the confidence level and hence helps the user for decision-making.

Recently, we introduced a wide panel of Mondrian CP models built using ChEMBL bioactivity datasets comprising nearly 600 human protein targets with various numbers of data points but also different ratios of active to inactive compounds[2]. Each of them was modelled individually and QSAR models were trained in parallel.

Herein, we present the models' results and focus on how CP performs on potentially flawed datasets.

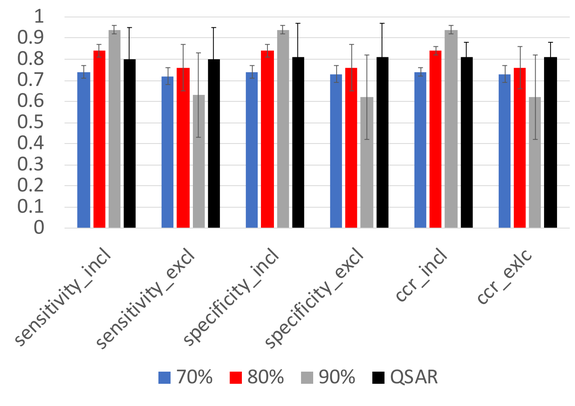
1- Model description

Datasets
- 550 human target bioactivity data from ChEMBL_23
- Between 76 and 7707 unique compounds associated (median 326)
- Ratios of active to inactive compounds between 0.04 and 13 (median 0.9)
- Active/inactive determined according to protein family thresholds

Molecular descriptors
- RDKit morgan fingerprint + MW + logP + HBD + HBA + RTB + TPSA

Classification models
- ***QSAR***: data split into training (80%) and test sets (20%)
- ***Mondrian CP***: training set (80%) is divided between a proper training (70%) and calibration sets (30%). Test set uses the remaining 20%
- Random Forest (RF) with 100 repetitions with different splits for each but identical for both methods to use the same test sets
- RF with class weights and Mondrian CP (MCP) allow the models to deal with imbalanced data

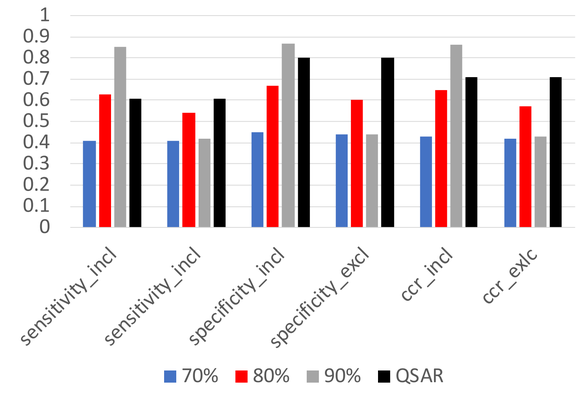
2- Internal validation



CCR = (sensitivity + specificity) / 2

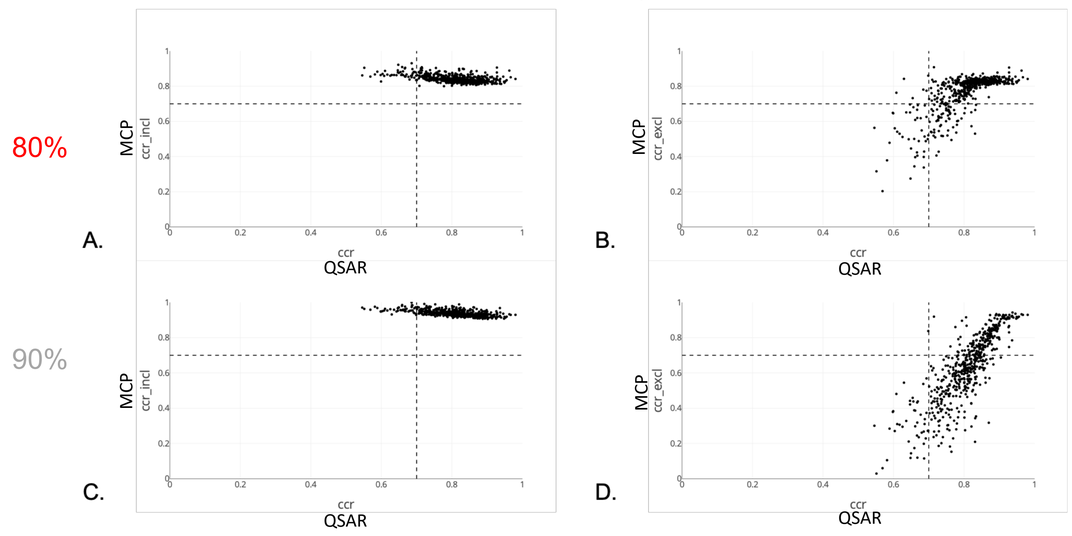
- MCP performance assessed at 70, 80 and 90% confidence levels
- MCP assigned predictions in 'active', 'inactive', 'both' or 'empty' classes
- 'empty' is always incorrect
- 'both' can be interpreted as informative and so considered as correct (*X_incl*), but it can also be seen as uninformative (*X_excl*)
- Good performance overall for MCP and QSAR models
- MCP performance affected at 90% confidence when 'both' class excluded

3- Temporal validation



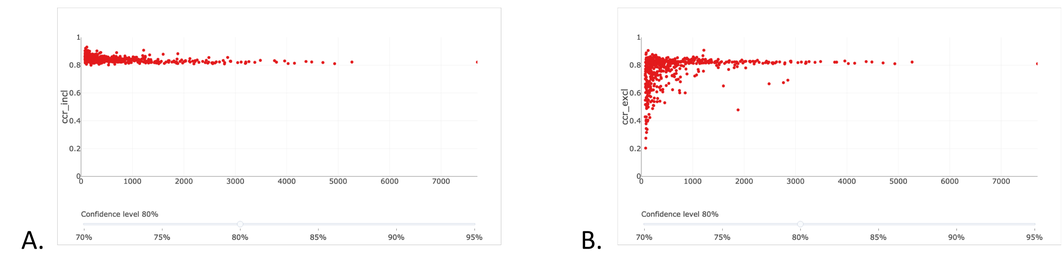
- Models were validated with data from ChEMBL_24, where available (515 targets)
- Sensitivity more affected but remains acceptable
- Difference between whether or not the 'both' class is considered shows MCP models have more difficulty in distinguishing actives and inactives

4- MCP vs QSAR

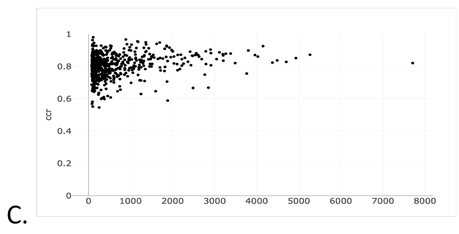


- When 'both' predictions considered as correct (A and C), MCP more likely to produce a model with CCR > 0.7 even with high confidence (80% and 90% here)
- When 'both' predictions are excluded (B and D), MCP and QSAR present similar results

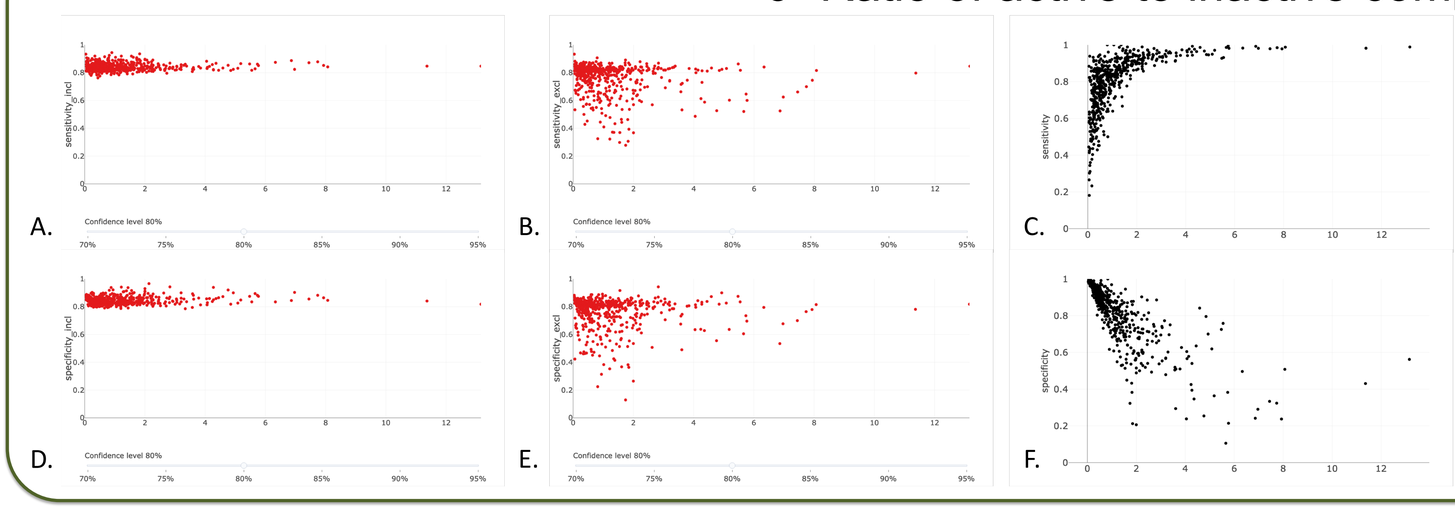
5- Number of data points



- Model panel presents a heterogenous number of data points
- With 'both' predictions considered as correct, MCP insensitive to the number of data points (A)
- Excluding 'both' predictions shows the impact of the lack of data on the MCP model performance (B)
- QSAR models are relatively little affected and maintain good performance even with few data points (C)



6- Ratio of active to inactive compounds



- MCP and QSAR models are affected differently by the ratio of active to inactive compounds
- With 'both' predictions considered as correct, MCP insensitive to the ratio no matter which class is overrepresented (A and D)
- Excluding 'both' predictions has an impact which is not clearly ratio dependent (B and E). Here MCP helps in maintaining good predictivity even for the minority class
- QSAR models are clearly affected by the ratio with those with an overrepresented active class exposing systematically higher sensitivity (C) and those with more inactives having a higher specificity (F)

Conclusions

1. QSAR models return good performance but lack confidence values and are affected by data distribution
2. MCP models return similar or higher performance at high confidence (80%) BUT…
3. … high confidence (90% and higher) implies less certainty in the predictions
4. No effect of the size of the dataset or of the class imbalance if 'both' predictions considered as correct
5. Ignoring these predictions degrades the performance which then become similar to QSAR models

(1) Norinder, U. et al. *J. Chem. Inf. Model.* **2014**, *54* (6), 1596–1603.
(2) Bosc, N. et al. *J. Cheminformatics* **2019**, *11* (1).

wellcome trust Strategic Award

EMBL-EBI

"This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No777365 ("eTRANSAFE"). This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA."

EMBL-EBI
Wellcome Trust Genome Campus
Hinxton, Cambridgeshire, CB10 1SD, UK

www.ebi.ac.uk/chembl/