XXXII Symposium on Bioinformatics and Computer-Aided Drug Discovery

Assay-Aware Learning-to-Rank for Robust Molecular Activity Prediction from Heterogeneous Bioactivity Data

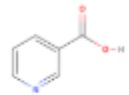
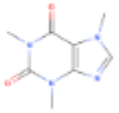
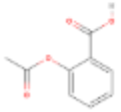
Luis Felipe de Morais Melo

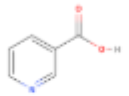
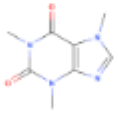
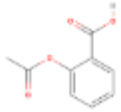
Federal University of Paraiba, Brazil

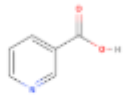
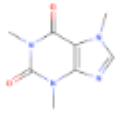
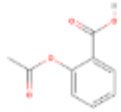
CHEMINFORMATICS LAB

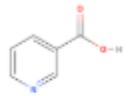
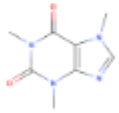
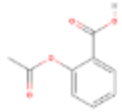
Introduction

- What is heterogeneous bioactivity data?

Laboratory 1	
Molecule	Activity
	6.5
	5.9
	5.5

Laboratory 2	
Molecule	Activity
	6.8
	6.2
	5.8

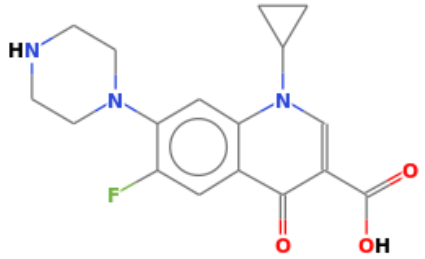
Laboratory 3	
Molecule	Activity
	6.0
	5.4
	5.0

Laboratory 4	
Molecule	Activity
	7.5
	6.9
	6.5

Introduction

- Are the results from the different labs **comparable** with eachother?

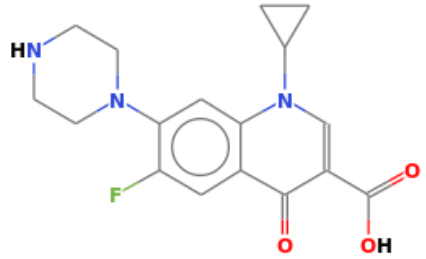
Laboratory 1



?

=

Laboratory 2



Introduction

- Are the results from the different labs **comparable** with eachother?

- **65%** of overlapping data points differ by **> 0.3 log units**;
- **27%** of data points differ by **> 1 log unit** (a 10-fold difference in potency).



Introduction

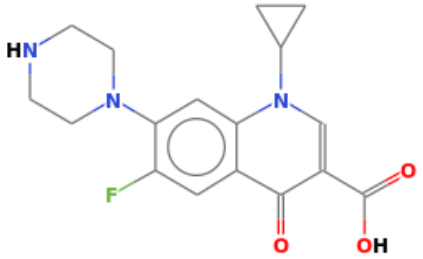
- Does the **relative ranking** of compounds **within an assay** provide a more consistent learning signal than absolute activity values?

Introduction

- Does the **relative ranking** of compounds **within an assay** provide a more robust and consistent learning signal for predictive modeling than absolute activity values?

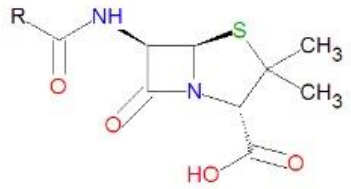
Laboratory 1

Molecule A



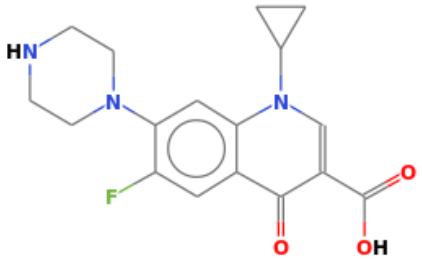
>

Molecule B



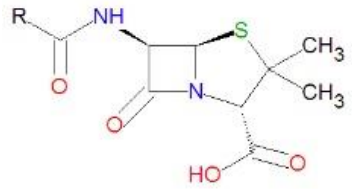
Laboratory 2

Molecule A



?

Molecule B



Introduction

- Does the **relative ranking** of compounds **within an assay** provide a more robust and consistent learning signal for predictive modeling than absolute activity values?

- If the answer to the above is **yes**, can a model trained to exploit this **relative ranking** perform better than a traditional QSAR model?

Materials and methods

- We selected all datapoints from **ChEMBL** where:
  - Targets were **individual proteins**;
  - The assay was a **binding** assay;
  - The assay metric was **IC50**;
  - The assay units were in **nM**;
  - The target had at least **3 distinct assays**;
  - The assay had at least **10 distinct molecules**.



Materials and methods

- For all of the obtained target datasets we:
  - Generated all potential assay combinations within the same target;
  - If there were at least a pair of duplicate molecules across the assays, the consistency of the internal ranking between the assays was calculated;

Materials and methods

- For all of the obtained target datasets:
  - We trained and evaluated (LOAO) 3 distinct models:
    - Baseline Regressor
    - Heterogenous model (assay-aware LambdaMART model)
    - Homogeneous model (assay-unaware LambdaMART model)

dmlc XGBoost

Materials and methods

- For all of the obtained target datasets:
  - We trained and evaluated (LOAO) 3 distinct models:
    - Baseline Regressor
    - Heterogenous model (assay-aware LTR model)
    - Homogeneous model (assay-unaware LTR model)
  - For each model we calculated the:
    - Pairwise accuracy
    - Spearman's rank correlation coefficient
    - Kendall's Tau
  - Finally, we compared the results of the Baseline regressor the Heterogenous model and the Homogeneous model.

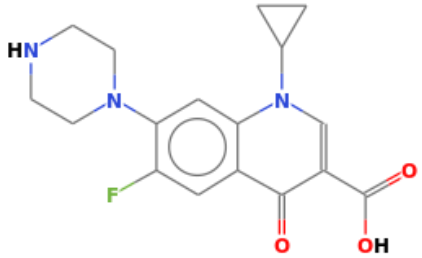
dmlc XGBoost

Results and Discussion

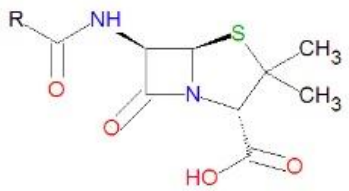
- After all the filtering and cleaning procedures, we obtained **1097** targets;

- The relationship between pairs of molecules on different assays was **consistent in 86%** of the analysed duplicates;

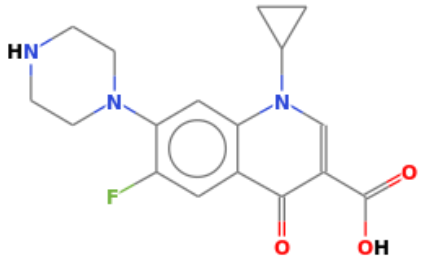
Laboratory 1



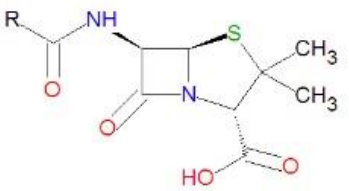
>



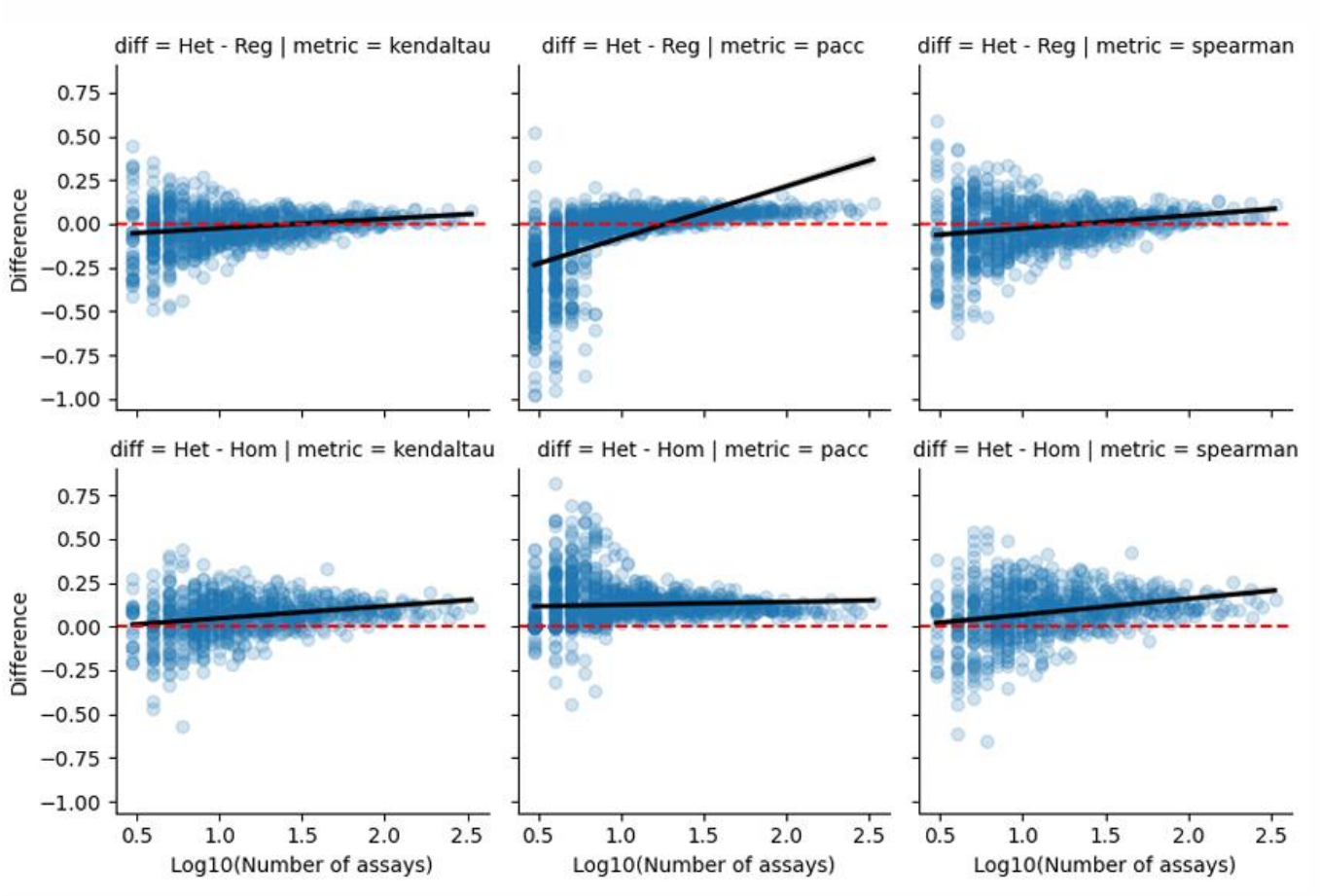
Laboratory 2



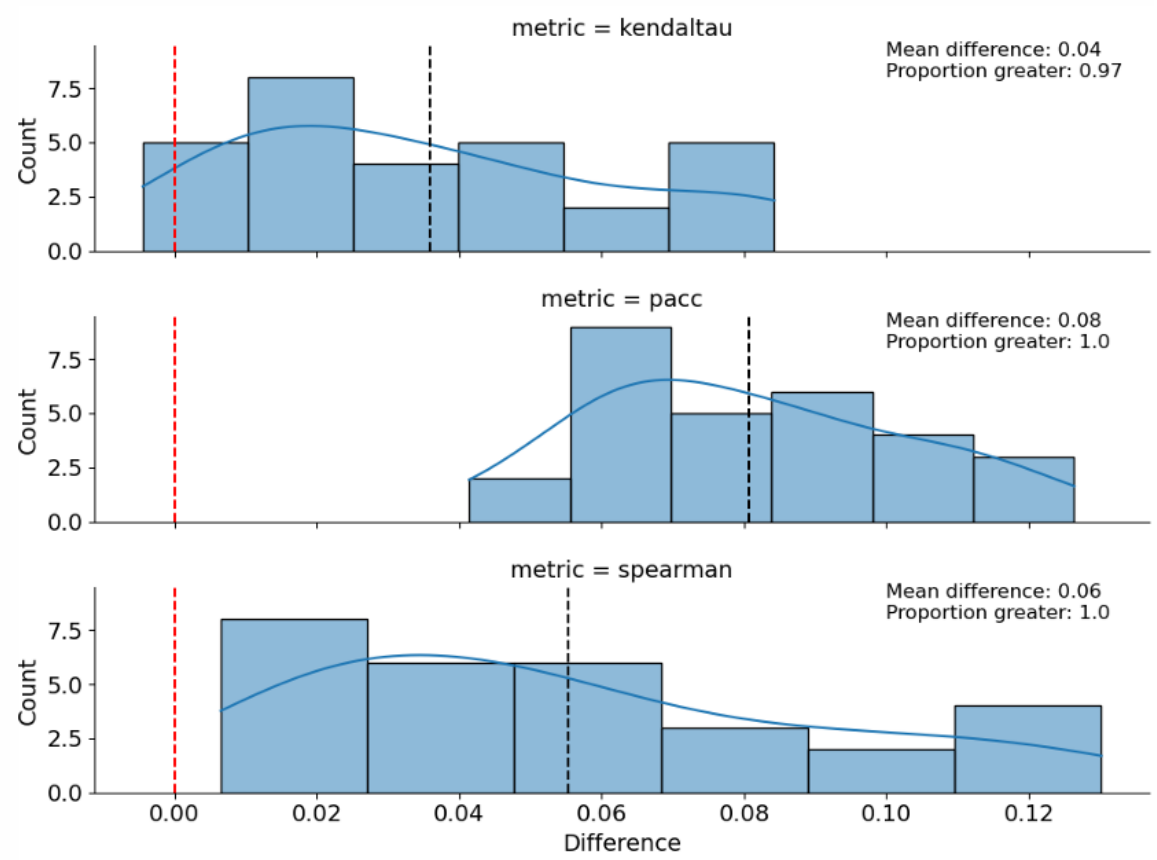
86%



Results and Discussion



Results and Discussion



Conclusion

- Relative activity relationships can provide a useful learning signal for heterogeneous pharmacological datasets.

- Explicitly accounting for assay boundaries can improve generalization to unseen assays, particularly when sufficient assay diversity is available.

- Although consistent, the observed improvements were moderate, and further methodological developments are warranted.

Thank you!

Luis Felipe de Morais Melo

luisfelipe.melo@hotmail.com

CHEMINFORMATICS LAB