Conformational Oversampling as Data Augmentation for Molecules

Jennifer Hemmerich[1*] , Ece Asilar[1] , Guenter Klambauer[2] and Gerhard F. Ecker[1]

[1]Department of Pharmaceutical Chemistry, University of Vienna, 1090 Vienna, Austria
[2]Institute of Bioinformatics, Johannes Kepler University Linz, Austria

*Contact: jennifer.hemmerich@univie.ac.at

Idea

Neural Networks have shown to be a viable tool for prediction of bioactivities. However, bioactivity datasets tend to be very small and imbalanced.

The problems:

(i) Neural networks overfit easily with small datasets.
(ii) Neural networks tend to ignore the minority class for imbalanced datasets.

Proposed solution:

(i) JU Bioactivities: a python package with toxicity data comprising 119188 compound from 476 non-unique endpoints
(ii) Augmentation helps neural network training for image recognition, can we use that for molecules?

JUBioactivities library

Sources — 9 webpages — 5 publications

Python package

119188 compounds (1212 invalid structures)

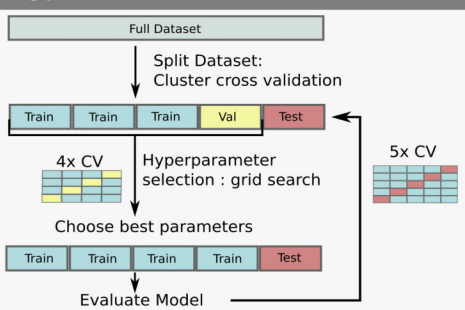
476 Endpoints non-unique

Binary Tasks:	76796 Molecules	184 Endpoints
Regression Tasks:	42998 Molecules	288 Endpoints
Multi-class Tasks:	29391 Molecules	4 Endpoints

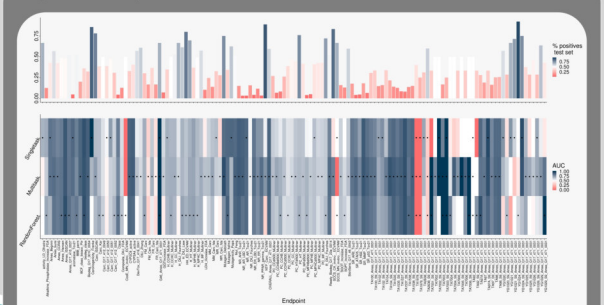
JUBioactivities example use

The library enables the user to easily read in available datasets and join them using a generated InChIKey. To ensure standardised identifiers the library uses structure standardisation[1] and then generates the InChIKey with RDKit[2]

Training procedure



JUNet comparison to classification models



The mutlitask network JUNet outperforms single task models, as well as, random forests in ~60% of the endpoints using the binary datasets from JUBioactivities, the black dots represent the best model per endpoint

Conformational oversampling (COVER) Workflow

Initial dataset

Tox21 dataset[1] Endpoint: p53 activation ("SR-p53")
External evaluation: Test sets from the challenge

	Test	External
Actives	371	56
Inactives	5741	677
Imbalance ratio	1:16	1:13

Data curation

Structure Standardisation:
removal of clashing activities and removal of mixtures and duplicates

Conformer generation

calculate imbalance ratio

$r = \left\lfloor \frac{n_{maj}}{n_{min}} \right\rfloor$ n_{maj} = no. of majority class compounds, n_{min} = no. of minority class compounds

minority class → majority class

RDKit ETKDG

m*r conformations — m conformations

no conformational selection

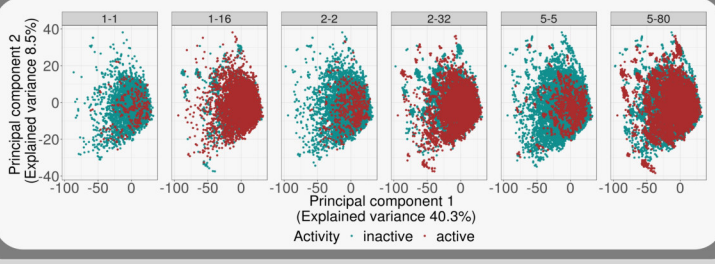
Training datasets

dataset	no. of conformations per inactive	no. of conformations per active	no. of molecules inactive	no. of molecules active	no. of molecules overall
1-1	1	1	5502	341	5843
1-16	1	16	5502	5428	10930
2-2	2	2	11001	680	11681
2-32	2	32	11001	10865	21866
5-5	5	5	27504	1698	29202
5-80	5	80	27504	27145	54649

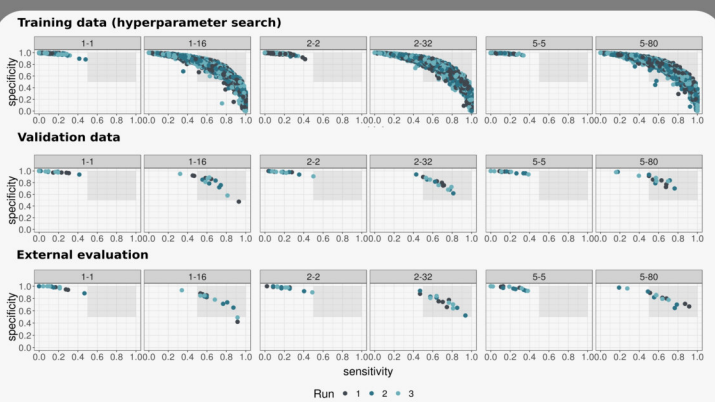
3D descriptors

117 MOE i3D descriptors
1028 DRAGON 7 3D descriptors

PCA of the oversampled datasets



Model performances



COVER increases the sensitivity of the models, leading to models which are able to classifiy molecules from both classes

Conclusion

Did we find solutions?

(i) Yes, JUBioactivities could be used to train a multitask network, **outperforming** other single task approaches in **60%** of the cases.

(ii) Yes, COVER helps to increase the training performance, especially with respect to sensitivity. However, oversampling alone does not increase the performance, **balancing is necessary** to increase the performance of the models.

Acknowledgements

This project has received funding from the Innovative Medicines Initiative 2 JointUndertaking under grant agreement No 777365 ("eTRANSAFE"). This Joint Undertaking receives support from the European Union's Horizon 2020 research and innovation programme and EFPIA. It also has received funding from the Austrian Science Fund FWF (grant W1232).

References

[1]https://tripod.nih.gov/tox21/challenge/