ICANN19

28th International Conference on Artificial Neural Networks

Conformational Oversampling as Data Augmentation for Molecules

Jennifer Hemmerich[1*] , Ece Asilar[1] , and Gerhard F. Ecker[1]
[1]University of Vienna, 1090 Vienna, 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:

Augmentation helps neural network training for image recognition, can we use that for molecules?

Methods

COVER Workflow

Initial dataset

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

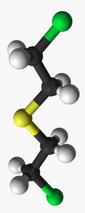
	Test	External
Actives	371	56
Inactives	5741	677
Imbalance ratio	1:16	1:13

Data curation

Structure Standardisation

removal of clashing activities

Conformer generation



calculate imbalance ratio

$$r = \left\lfloor \frac{n_{maj}}{n_{min}} \right\rceil$$

n_{maj} = no. of all compounds
n_{min} = no. of minority class compounds

minority class — majority class

RDKit ETKDG

m*r conformations — m conformations

no conformational selection

Training datasets

	no. of conformations per		no. of molecules		
dataset	inactive	active	inactive	active	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

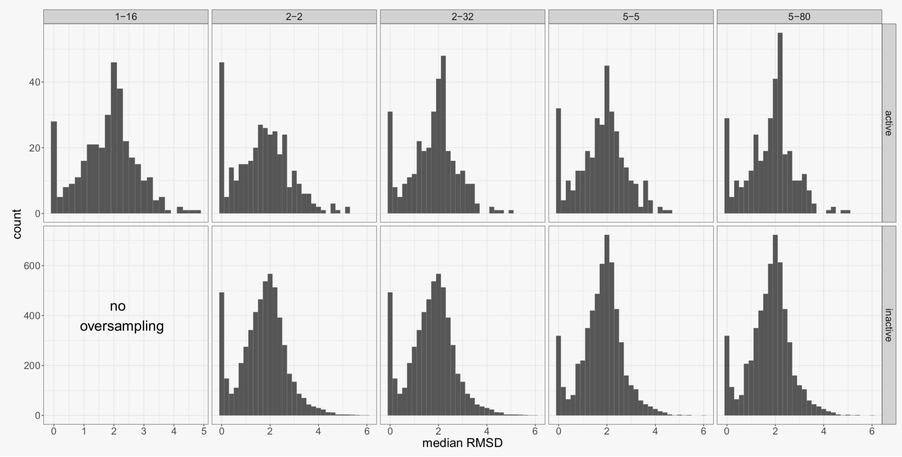
3D descriptors

117 MOE i3D descriptors
1028 DRAGON 7 3D descriptors

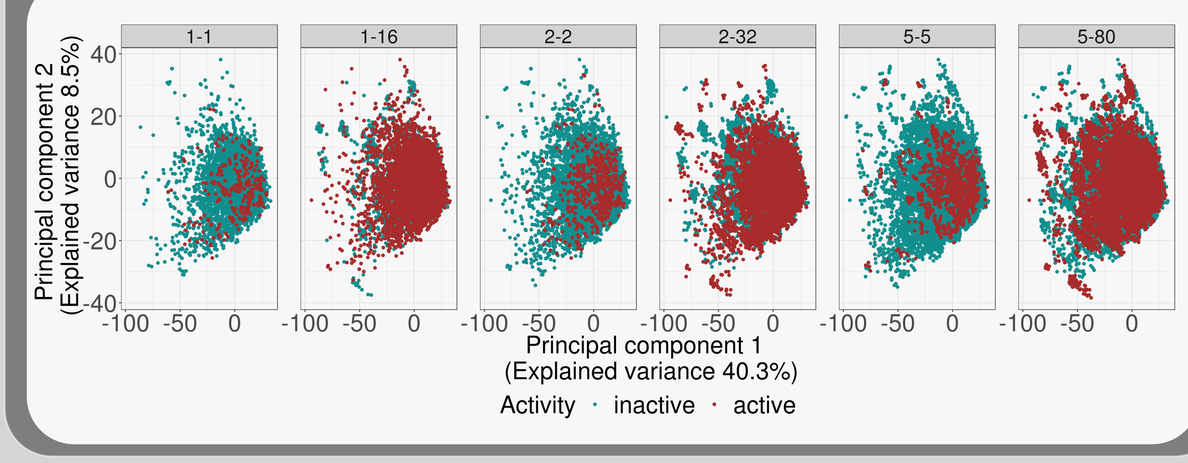
Training procedure

Results

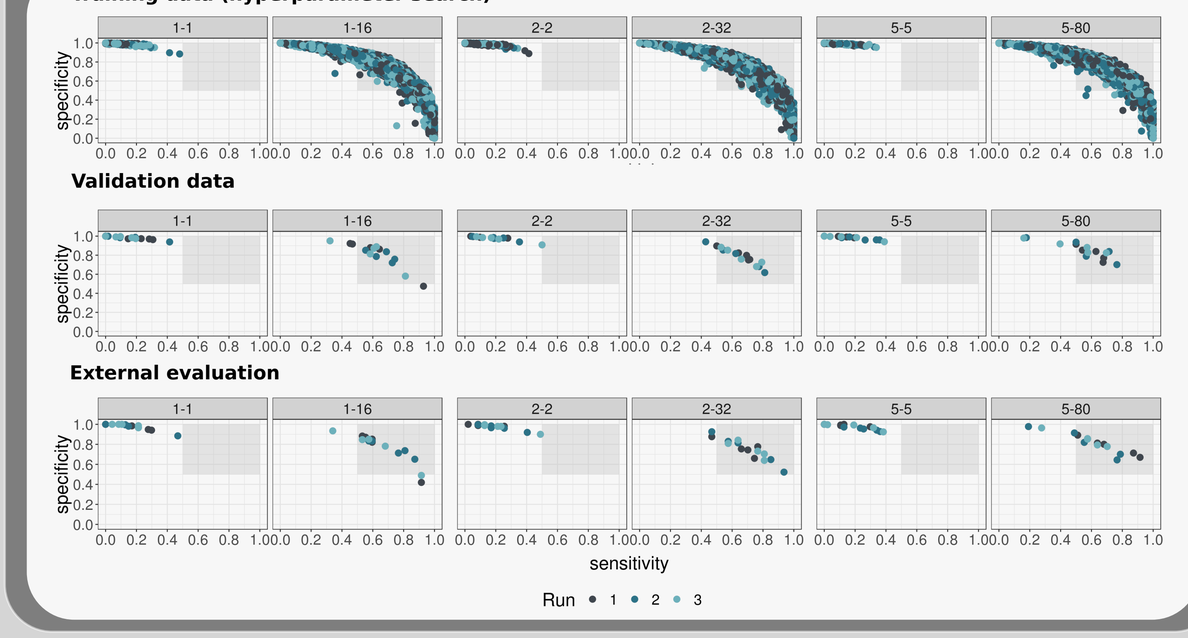
Comparison of median RMSDS



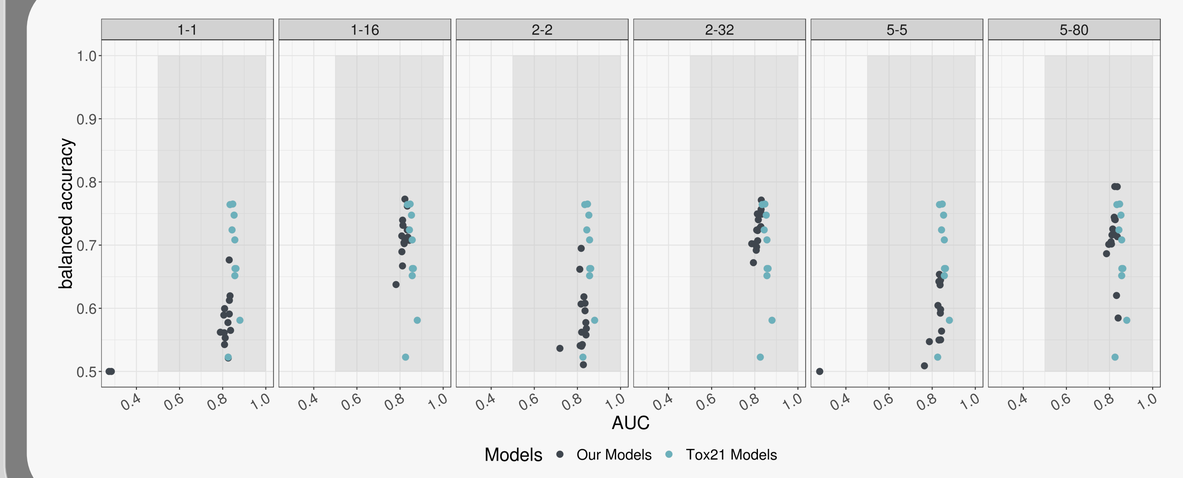
PCA of the oversampled datasets



Model performances



Comparison to Tox21 top 10 models



Conclusion

Did we find a solution?

Yes, COVER helps to increase the training performance, especially with respect to sensitivity

But:

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/