Conformational Oversampling as Data Augmentation for Molecules

Jennifer Hemmerich^{ID}, Ece Asilar^{ID}, and Gerhard F. Ecker^(✉)^{ID}

University of Vienna, 1090 Vienna, Austria
{jennifer.hemmerich,ece.asilar,gerhard.f.ecker}@univie.ac.at

Abstract. Toxicological datasets tend to be small and imbalanced. This quickly causes models to overfit and disregard the minority class. To solve this issue we generate conformations of molecules. Thereby, we can balance datasets as well as increase their size. Using this approach on the Tox21 Challenge data we observed conformational oversampling to be a viable approach to train datasets, increasing the balanced accuracy of trained models.

Keywords: Deep learning · Toxicity · Chemoinformatics · Imbalanced learning

1 Introduction

Safety evaluation during drug development would greatly benefit from models reliably predicting specific toxicities. Such models could reduce time and cost as well as animal testing by indicating hazards in the first stage of drug design. The Merck Kaggle competition and the Tox21 Challenge demonstrated the superiority of neural networks over traditional machine learning approaches for biological activity predictions [7, 8, 11, 12]. Hence, neural networks are a viable tool to generate predictive models for toxicities. However, toxicological datasets often are small and imbalanced. This oftentimes results in overfitting of models or neglect of the minority class, predicting almost all compounds as the majority class. Especially for toxicity predictions, it is important that models exhibit high sensitivity and specificity, that is, identify hazards but not flagging every single compound.

With the aim to solve both, imbalance and overfitting, we developed a method named COVER (Conformational OVERsampling). This method was derived from augmentation as used in image recognition, where images are transformed to increase generalization and dataset size (e.g. [3, 5, 10, 13]). Instead of transforming an image we “transform” molecule representations by generating multiple conformations along with a 3D based description of molecules. This enables us to balance datasets for neural network training, as well as increase the dataset size.

Table 1. Datasets used for training with the number of molecules per class and the overall dataset size, each conformation is counted as separate molecule

Dataset	No. of conformations per		No. of molecules		
	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

Fig. 1. Sensitivity versus specificity of the trained models. Each datapoint presents the evaluation of one model. The captions of the subplots denote the amount of oversampling done, e.g. 1_16 denotes 1 conformation for the negative class and 16 conformations for the positive class. Each shape with the respective color denotes an independent run of the cross-validation scheme. The area in the upper right corner shows the region of predictive models (model with a sensitivity and specificity higher than 0.5). (A) Plot of the models obtained during hyperparameter selection, (B) performance of the final models on the external fold, (C) performance of the final models on the external test set. (Color figure online)

2 Method

For model training we assembled a dataset using the the endpoint p53 activation (“SR-p53”) from the Tox21 Challenge data which contains 5843 compounds with binary annotation. The endpoint has an imbalance ratio of 1:16, denoting that the dataset contains 16 times more inactive than active molecules. In the next

step we generated conformations of the molecules using the ETKDG Algorithm with energy minimisation as implemented in RDKit [6,9]. In total, we generated 6 datasets (see Table 1). For each dataset m conformations were generated for the inactives and n conformations for the actives, resulting in the label “ m - n dataset”. Therefore, to generate m conformations the conformation generation was run m times, without any further processing. The datasets were prepared with different goals: Firstly to evaluate the training on the original dataset (“1-1 dataset”), secondly to evaluate training on a balanced dataset (“1-16 dataset”), thirdly to evaluate whether oversampling alone is beneficial (“2-2” and “5-5 dataset”) or if balancing is mandatory (“2-32” and “5-80 dataset”). After conformer generation 1145 3D descriptors were calculated, which are used as input for the models. In addition, we used the two test datasets from the Tox21 Challenge as a combined external validation dataset containing 643 compounds.

To erase increased predictivity by sophisticated network architectures we only trained multilayer feed-forward neural networks with two to four hidden layers. The training was conducted using a nested cross-validation [1]. The inner cross-validation is used for hyperparameter grid search, whereas the outer loop is used to validate the model on an external validation fold. In addition, the splits of the dataset are not chosen randomly but decided via affinity propagation clustering [4,7]. The generated clusters are randomly distributed to the five folds, such that molecules of the same cluster are found in the same fold. The thereby achieved low similarity between folds ensures a reduction in the model bias. To prevent bias by the increased number of conformations of single molecules, the conformer generation was done after the clustering and splitting. Subsequently, each conformation was assigned the same cluster number, and thus cross validation fold, as the parent molecule. This prevents a leakage of information between the training and test dataset and the different cross validation folds.

Since the focus of model training was the suitability for toxicity predictions, balanced accuracy was used to choose the best hyperparameter set. Balanced accuracy is calculated as the harmonic mean between sensitivity and specificity.

3 Results and Discussion

For all datasets we could generate the necessary number of conformations. For rigid molecules, which constituted about 10% of the dataset, we obtained very similar or the same conformation. Nevertheless, the conformers of one molecule had an average root mean square deviation (RMSD) of about 1.7–1.9 Å.

Using the dataset with one conformation per molecule (1-1 dataset) we observed that training yielded good results with respect to the area under the receiver operating curve (AUC) but, in most cases, balanced accuracy was lower than 0.6. In these cases, the models have a high specificity but lack sensitivity. Evaluating the predictions showed that mostly all molecules were classified as the majority class. Applying COVER, we observed considerable changes. Specifically, trained models gained sensitivity with only slight loss of specificity and no loss in the AUC. This indicates the models are able to predict both classes.

Figure 1A illustrates oversampling by itself does not increase predictivity of models trained with the 2–2 or 5–5 dataset. Contrarily, models trained with the 1–16, 2–32 and 5–80 datasets show an increase in the sensitivity, and thus predictivity. The results for the external fold of the cross-validation can be seen in Fig. 1B. It shows models trained with balanced datasets do not suffer from low sensitivity. To ascertain the models also work for an external dataset we used the test set from the Tox21 challenge. No model showed a decrease in predictivity, which is depicted in Fig. 1C.

Our approach demonstrates creation of multiple conformations of a molecule facilitates training of neural networks. This is achieved using only information inherent in the dataset without having to create artificial samples as is often done in traditional machine learning (e.g. SMOTE [2]). Since our idea originated from image augmentation, we do not view the conformations as biological relevant, therefore all conformation have the same label as the parent molecule. Rather we hypothesize that through increase in training space the model is able to generalize better. This is similar to augmented images helping a network in the generalization. Therefore, rather than regarding the images as meaningful representations, they present the variety of the real world to a model. Thus no conformation selection process is implemented. In the future it would be very interesting to investigate the influence of a more sophisticated conformer selection on model training. However, this makes balancing harder as there might not be enough distinct conformations per molecule.

In general, we observed COVER is only beneficial when also balancing the dataset. By oversampling, the training space of the network is increased, facilitating training. Subsequently, models cannot disregard the minority class. Our final validation showed that models trained on datasets balanced with COVER have a much higher sensitivity than those trained on unbalanced datasets.

Acknowledgements. This project has received funding from the Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 777365 (“eTRANSafe”). This Joint Undertaking receives support from the European Union’s Horizon 2020 research and innovation programme and EFPIA.

References

1. Baumann, D., Baumann, K.: Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *J. cheminf.* **6**(1), 47 (2014). <https://doi.org/10.1186/s13321-014-0047-1>
2. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002). <https://doi.org/10.1613/jair.953>
3. Ciresan, D.C., Meier, U., Gambardella, L.M., Schmidhuber, J.: Deep big simple neural nets excel on handwritten digit recognition. *Neural Comput.* **22**(12), 3207–3220 (2010). https://doi.org/10.1162/NECO_a.00052
4. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. *Science* **315**(5814), 972–976 (2007). <https://doi.org/10.1126/science.1136800>

5. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q. (eds.) *Advances in Neural Information Processing Systems* 25, pp. 1097–1105. Curran Associates, Inc., New York (2012). <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
6. Landrum, G.: RDKit: open-source cheminformatics (2006). <http://www.rdkit.org/>
7. Mayr, A., Klambauer, G., Unterthiner, T., Hochreiter, S.: DeepTox: toxicity prediction using deep learning. *Front. Environ. Sci.* **3**, 80 (2016). <https://doi.org/10.3389/fenvs.2015.00080>
8. MerckKaggle: Merck Molecular Activity Challenge (2012). <https://www.kaggle.com/c/MerckActivity>
9. Riniker, S., Landrum, G.A.: Better informed distance geometry: using what we know to improve conformation generation. *J. Chem. Inf. Model.* **55**(12), 2562–2574 (2015). <https://doi.org/10.1021/acs.jcim.5b00654>
10. Simard, P.Y., Steinkraus, D., Platt, J.C.: Best practices for convolutional neural networks applied to visual document analysis. In: *Proceedings of the Seventh International Conference on Document Analysis and Recognition 2003*, pp. 958–963, August 2003. <https://doi.org/10.1109/ICDAR.2003.1227801>
11. Team, K.: Deep Learning How I Did It: Merck 1st place interview, November 2012. <http://blog.kaggle.com/2012/11/01/deep-learning-how-i-did-it-merck-1st-place-interview/>
12. Tox21: Tox21 Data Challenge 2014 (2014). <https://tripod.nih.gov/tox21/challenge/>
13. Wong, S.C., Gatt, A., Stamatescu, V., McDonnell, M.D.: Understanding data augmentation for classification: when to warp? In: *2016 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp. 1–6, November 2016. <https://doi.org/10.1109/DICTA.2016.7797091>

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.