








Augmentation Is What You Need!

Igor V. Tetko¹ , Pavel Karpov¹ , Eric Bruno² ,
Talia B. Kimber³ , and Guillaume Godin³ 

¹ Institute of Structural Biology, Helmholtz Zentrum Muenchen -
German Research Center for Environmental Health (GmbH) and BIGCHEM
GmbH, Neuherberg, Germany

{i.tetko, pavel.karpov}@helmholtz-muenchen.de

² Expedia, Geneva, Switzerland
ebruno@expedia.com

³ Research and Development Division, Firmenich International SA,
Geneva, Switzerland

talia.kimber@gmail.com, guillaume.godin@firmenich.com

Abstract. We investigate the effect of augmentation of SMILES to increase the performance of convolutional neural network models by extending the results of our previous study [1] to new methods and augmentation scenarios. We demonstrate that augmentation significantly increases performance and this effect is consistent across investigated methods. The convolutional neural network models developed with augmented data on average provided better performances compared to those developed using calculated molecular descriptors for both regression and classification tasks.

Keywords: Augmentation · Convolutional neural networks · Descriptor representation · QSAR · Chemoinformatics · Regression · Classification

1 Introduction

The renaissance of neural networks methods [2, 3] has brought a wide variety of neural network types including methods that are capable to analyze chemical structure represented as graphs or text (e.g., SMILES) in chemistry. These methods are particularly interesting since their internal representations, latent variables, in principle do not lose information about the chemical structures as compared to methods using chemical descriptors. These latent variables can be used to decode the chemical structures and address the problem of inverse Quantitative Structure Activity Relationship (QSAR) [4], which has been a challenge for chemoinformatics since first QSAR models. However, the practical question arises whether the prediction power of such methods is similar to those based on descriptors. In our previous study [1] we have demonstrated that the Convolutional Neural Fingerprint (CNF) which is based on ideas of text processing originally proposed by [5], provided similar performance to models developed using three descriptors sets. The high prediction accuracy of the CNF was achieved thanks to the augmentation technique, which was originally proposed in

computer vision and was recently introduced to QSAR studies [6]. It is worth mentioning that a related technique to enhance accuracy of QSAR models by considering symmetry of molecules was proposed more than 20 years ago [7]. Typically so called canonical SMILES produced according to some rule-defined enumeration of atoms, are used to train the model. The augmentation procedure generates a number of unique SMILES for the same molecule, e.g., by starting enumeration of atoms from a random atom and/or traversing molecular graph path in random order. Augmentation is employed during both training and inference steps: during training, augmentation increases the dataset size by providing for each structure a number n of distinct SMILES; during inference, the prediction for a given structure is averaged over predictions of m distinct SMILES of that structure.

The goal of this study was to clarify whether the performance of augmented models is similar to those developed using a large set of descriptors typically used in QSAR studies. We also include a result for TextCNN [8] which is a DeepChem implementation of the Char-CNN [5], but using a different architecture.

2 Methods

The CNF method from our previous study [1] as well as the TextCNN method as implemented in DeepChem [8] were used as convolutional methods. Associative Neural Networks (ASNN) [9], which is a shallow neural network was used as a traditional method to develop models using descriptors. Early stopping was used to prevent overfitting of neural networks [10] for all three analyzed methods.

Augmentation. The augmentation used for convolutional methods was generated with RDKit and included: (1) **no-augmentation** – canonical SMILES was used, (2) **off-line augmentation** with $n = 10$ SMILES generated before the neural network training and (3) **on-line augmentation** in which new SMILES were generated for each training epoch (only for CNF). The augmentation with $n = 10$ SMILES (which was selected based on results in [1]) was also applied during the prediction step and the average value was used as the final model prediction. It was shown that model performance decreased when model developed with canonical SMILES was used to predict augmented data [1]. Therefore for models developed with the first protocol only canonical SMILES were used during the prediction step.

Hyperparameter Optimization. The methods were used with their default parameters as available on the On-line Chemical Database and Modeling Environment (<http://ochem.eu>). For CNF we used the same parameters as in [1] with an exception of the convolutional filter, which was increased to 5. We tried to optimize neural network parameters, such as the number of neurons, architecture, activation function, etc. but did not see improvement as compared to the defaults. The full run of optimization required more than 12 h on six GPU cards (GeForce RTX 2070 and 1070, Quadro P6000, Titan Xp and V) and thus exhaustive investigation of all options was impossible. Since training with augmented data was about 10 times longer for each epoch, the number of epochs for on-line training was respectively increased 10 times to allow networks to use about the same number of training steps.

Descriptors. In total 16 sets of descriptors, namely ALogPS + OEstate, CDK2, ChemaxonDescriptors, Dragon7, Fragmentor, GSfrag, InductiveDescriptors, JPllogP, Mera + Mersy, PyDescriptor, QNPR, RDKit, SIRMS, Spectrophores, StructuralAlerts and alvaDesc were used with their default settings. The descriptors are described on the OCHEM web site and were used in multiple previous studies. Many of the descriptors have their own hyper parameters, e.g., size of fragments for fragmental descriptors, which can be also optimized. We did not perform such optimization for this study but instead used default hyperparameters that were found to be optimal ones in the previous studies.

Model Validation. Five-fold cross-validation was used to test performance of all models.

Datasets. The same 9 regression and 9 classification sets from our previous study [1] were used.

Statistical Parameters. The regression models were compared using the coefficient of determination

$$r^2 = 1 - \frac{\sum (f_i - y_i)^2}{\sum (\bar{y} - y_i)^2}$$

where \bar{y} is the average value across all samples, while f_i and y_i are predicted and target values for sample i , respectively. The classification results were compared using Area Under the Curve (AUC).

3 Results

The augmentation dramatically improved results for CNF method but also contributed better models for TextCNN for both regression and classification datasets as shown on Figs. 1 and 2.

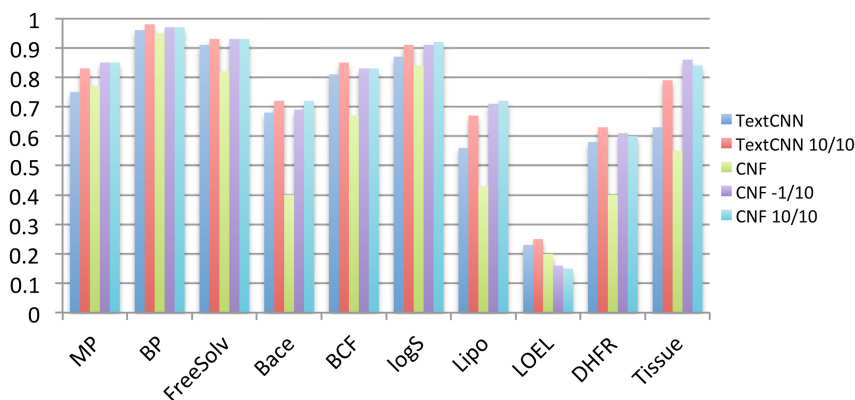


Fig. 1. Coefficient of determination, r^2 , for regression tasks. With an exception of LOEL dataset, which had the lowest r^2 , the training with augmentation improved the accuracy of models for all datasets. “10/10” and “-1/10” indicate off-line and on-line augmentations, respectively.

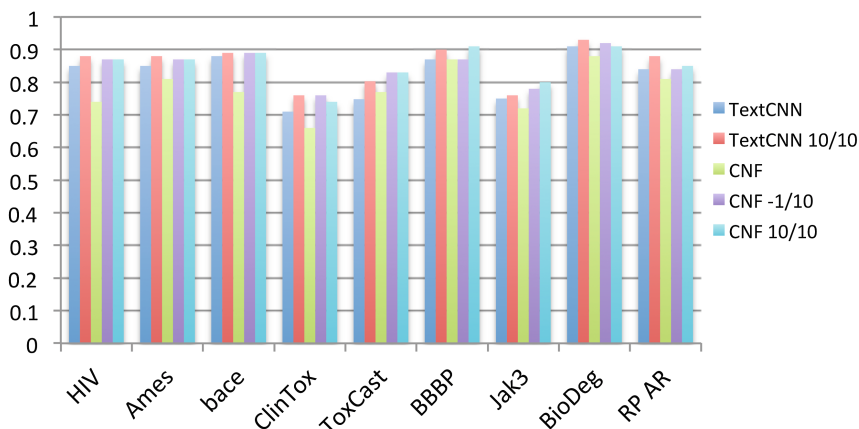


Fig. 2. For classification tasks the augmentation consistently improved AUC values for both convolutional neural networks.

For comparison of performances of models developed using descriptors and convolutional neural networks, we counted the number of models for which ASNN (using any descriptor set) or one of augmented convolutional models provided better results. Such comparison was biased towards ASNN since the best result for this method was selected from 16 models corresponding to the used descriptor sets versus only three (two off-line and one on-line) models for SMILES-based approaches. For three datasets the best models for both approaches had the same performance. For remaining data, the SMILES-based approaches contributed better models in 11 cases while descriptor-based approaches did it for 4 models.

4 Conclusions

We showed that convolutional neural networks trained with augmented data provide better performances compared to models developed with the state-of-the-art descriptor representation of molecules for both regression and classification problems.

Acknowledgements. This study has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 676434, "Big Data in Chemistry" and ERA-CVD "Cardio-Oncology" project, BMBF 01KL1710. The authors thank NVIDIA Corporation for donating Quadro P6000, Titan Xp and Titan V graphics cards for this research. We also thank ChemAxon (<http://www.chemaxon.com>) for Academic license of software tools as well as AlvaScience (<http://alvascience.com>), Molecular Networks GmbH (<http://mn-am.com>) and Chemosophia (<http://chemosophia.com>) for providing descriptors and Corina 2D to 3D conversion program used in this study.

References

1. Kimber, T.B., Engelke, S., Tetko, I.V., Bruno, E., Godin, G.: Synergy effect between convolutional neural networks and the multiplicity of SMILES for improvement of molecular prediction. eprint. [arXiv:1812.04439](https://arxiv.org/abs/1812.04439) (2018)
2. Baskin, I.I., Winkler, D., Tetko, I.V.: A renaissance of neural networks in drug discovery. *Expert Opin. Drug Discov.* **11**(8), 785–795 (2016). <https://doi.org/10.1080/17460441.2016.1201262>
3. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., Blaschke, T.: The rise of deep learning in drug discovery. *Drug Discov. Today* **23**(6), 1241–1250 (2018). <https://doi.org/10.1016/j.drudis.2018.01.039>
4. Blaschke, T., Olivecrona, M., Engkvist, O., Bajorath, J., Chen, H.: Application of generative autoencoder in De Novo molecular design. *Mol. Inform.* **37**(1–2), 1700123 (2018). <https://doi.org/10.1002/minf.201700123>
5. Zhang, X., LeCun, Y.: Text understanding from scratch. eprint. [arXiv:1502.01710](https://arxiv.org/abs/1502.01710) (2015)
6. Bjerrum, J.E.: SMILES enumeration as data augmentation for neural network modeling of molecules. eprint. [arXiv:1703.07076](https://arxiv.org/abs/1703.07076) (2017)
7. Baskin, I.I., Halberstam, N.M., Mukhina, T.V., Palyulin, V.A., Zefirov, N.S.: The learned symmetry concept in revealing quantitative structure-activity relationships with artificial neural networks. *SAR QSAR Environ. Res.* **12**(4), 401–416 (2001). <https://doi.org/10.1080/10629360108033247>
8. Ramsundar, B., Eastman, P., Walters, P., Pande, V.: *Deep Learning for the Life Sciences: Applying Deep Learning to Genomics, Microscopy, Drug Discovery, and More*. O'Reilly Media, Newton (2019)
9. Tetko, I.V.: Associative neural network. *Methods Mol. Biol.* **458**, 185–202 (2008). https://doi.org/10.1007/978-1-60327-101-1_10
10. Tetko, I.V., Livingstone, D.J., Luik, A.I.: Neural network studies. 1. Comparison of overfitting and overtraining. *J. Chem. Inf. Comput. Sci.* **35**(5), 826–833 (1995). <https://doi.org/10.1021/ci00027a006>

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.

