Synergy Effect between Convolutional Neural Networks and Multiplicity of SMILES

Talia B. Kimber[1], Sebastian Engelke[1], Igor V. Tetko[2], Eric Bruno[3], Guillaume Godin[3]

[1]University of Geneva, [2]Bigchem gmbh, [3]Firmenich SA

Motivation

The aim of this study[1] is to use a deep learning approach to predict molecular properties, such as the melting point of a chemical compound.

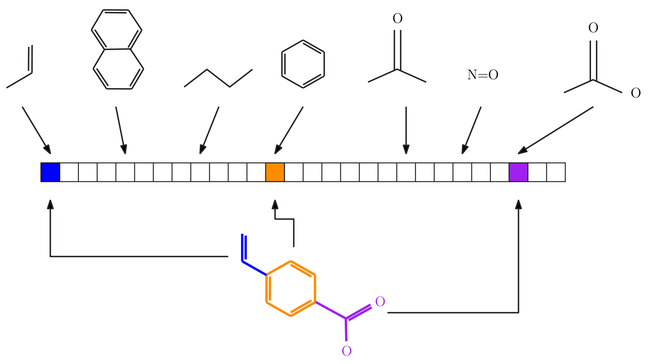
We use SMILES instead of the molecular graph[2] for computational efficiency. The convolution layers C in the CNF model extract neighbouring atoms in a chemical structure. The multiplication H is an embedding on a latent space. The procedure is similar to the Morgan fingerprint.

The CNF model maps a sparse matrix, the one-hot encoding of a SMILES, into a dense vector, the neural fingerprint. Two SMILES with similar spelling will be close, in the Euclidian sense, in the latent space.

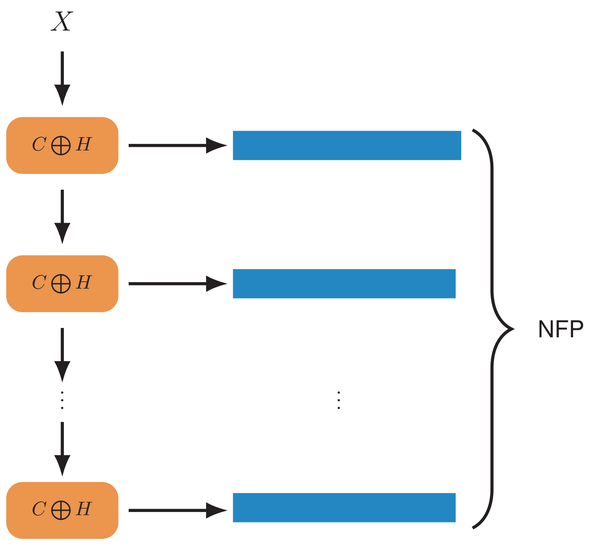
The NFP can then be plugged in other ML algorithms for molecular prediction.

Model

Fingerprint : a numerical representation which encodes molecular information



Convolutional Neural Fingerprint Model

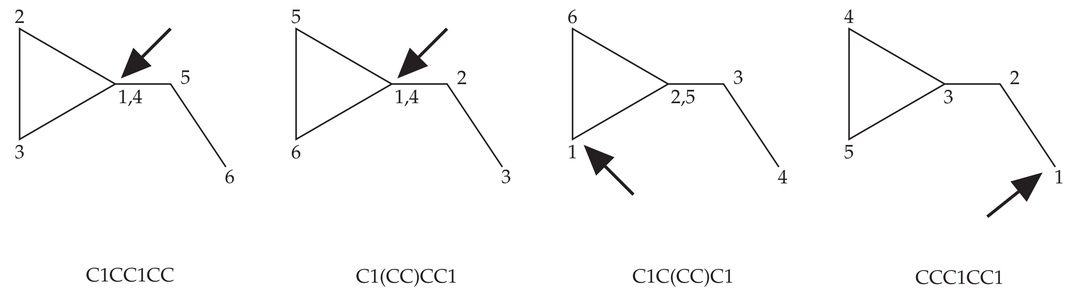


X: input to the CNF model; one-hot encoding of a SMILES
$C \bigoplus H$: convolutional plus a hashing layer
Pooling: sum by column
NFP: concatenation of vectors

Conclusion

SMILES augmentation improves the CNF model predictions by 25% in the best case. Moreover the CNF model generally performs better than highly fine-tuned models, which use traditional descriptors, such as RDKit, CDK2 and Dragon7.

Data Augmentation



SMILES : linearisation of a given molecular graph in a single line text, obtained by enumerating nodes and edges following a certain topological order

Multiplicity of SMILES : equally valid SMILES can be obtained for a single molecule [3], depending on:

- ◊ the atom where the enumeration starts
- ◊ the path followed along the 2D graph

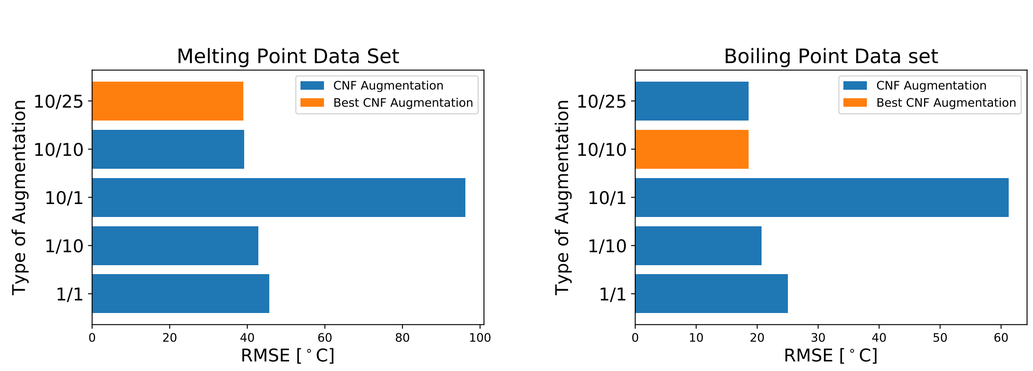
Problem : cycle, branch and stereochemistry breaks can create ambiguities when linearising a graph, especially for training a neural network

Idea : multiple SMILES exposes the neural network to different angles of the same object, changing the perspective from local to global

Type of Augmentation and Results

New SMILES are generated for each training epoch. The augmentation with $n = 10$ SMILES is also applied during the prediction step and the average value is used as the final model prediction [5].

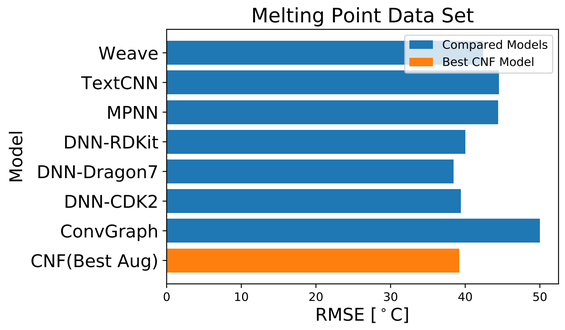
SMILES 1/1 No augmentation on SMILES
SMILES n/1 SMILES augmentation during training. *Always improves results*
SMILES 1/m SMILES augmentation during testing. *Almost always worsens results*
SMILES n/m SMILES augmentation during training and testing. *Best results*



The CNF model is applied on:

- ◊ regression and classification tasks (e.g. MP, BP, FreeSolv and HIV, Tox21, BioDeg)
- ◊ size of data sets: small (of order 10^2), medium (of order 10^3), large (of order 10^4)

Comparison with Other Models [4]



Weave Graph-based model
TextCNN Convolutional NN
MPNN Message Passing NN using LSTM
DNN-RDKit Deep NN using RDKit Descriptors
DNN-Dragon7 Deep NN using Dragon7 Descriptors
DNN-CDK2 Deep NN using CDK2 Descriptors
ConvGraph Graph-based model

References

[1] Kimber, T. B., Engelke, S., Tetko, I. V., Bruno, E., & Godin, G. (2018). Synergy Effect between Convolutional Neural Networks and the Multiplicity of SMILES for Improvement of Molecular Prediction. *arXiv preprint arXiv:1812.04439.*

[2] Duvenaud, D. K., Maclaurin, D., Iparraguirre, J., Bombarell, R., Hirzel, T., Aspuru-Guzik, A., & Adams, R. P. (2015). Convolutional networks on graphs for learning molecular fingerprints. In *Advances in neural information processing systems* (pp. 2224-2232).

[3] Bjerrum, E. J. (2017). Smiles enumeration as data augmentation for neural network modeling of molecules. *arXiv preprint arXiv:1703.07076.*

[4] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., ... & Pande, V. (2018). MoleculeNet: a benchmark for molecular machine learning. *Chemical science*, 9(2), 513-530.

[5] Tetko, I.V., Karpov, P., Bruno, E., Kimber, T.B., Godin, G. Augmentation is what you need! In *ICANN2019*, in press.