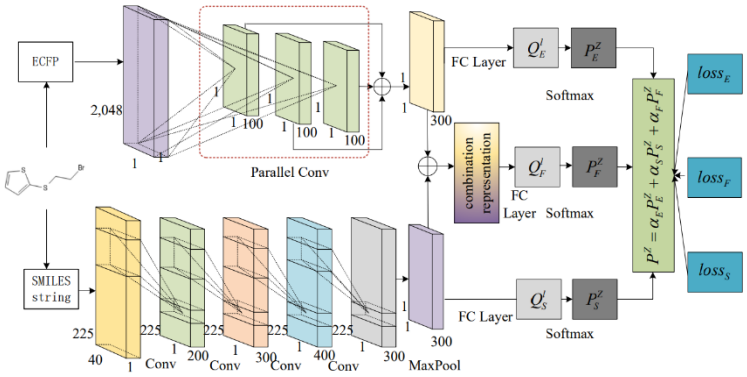
A Multi – Stream Network for Retrosynthesis Prediction

Qiang ZHANG, Juan LIU, Wen ZHANG, Feng YANG, Zhihui YANG, Xiaolei ZHANG

Frontiers of Computer Science, DOI: 10.1007/s11704-023-3103-z

Problems & Ideas

- Problems of existing retrosynthesis prediction approach:
  - The information represented by SMILES as well as the information depicted by ECFPs are both useful for retrosynthesis prediction.
  - Most existing methods only benefit from one kind of information rather than further consider the diverse aspect of molecular information.
- Ideas: A joint view to represent the molecules as both SMIELS strings and ECFPs to fuse the different kinds of information of the two descriptors.



The pipeline of MSNR: (i) The parallel CNN and the text-CNN that takes the ECFPs and SMILES after one-hot encoding as input to produce the deep features. (ii) The combination representation is devised by fusing the two kinds of deep features of ECFPs and SMILES. (iii) Three dense classifiers have been implemented, fusing these multi-stream prediction results with varying weights to arrive at a final retrosynthesis prediction, and the model is trained using an overall loss function.

Main Contributions

- Contributions:
  - MSNR jointly encodes the information of ECFPs, SMILES, and their combination to better and comprehensively describe the molecules, then MSNR makes the retrosynthesis prediction by weighting the multi-stream prediction results based on the above descriptors;
  - MSNR trained by a weighted loss function to further improve the stability and accuracy of retrosynthesis prediction;
  - The method achieves more competitive performance in retrosynthesis prediction than the compared methods under different top-k accuracies on different datasets.

Table 1 Top-k exact match accuracy in USPTO-50k dataset.

Method	Top-1	Top-3	Top-5	Top-10
RetroSim [1]	37.3	54.7	63.3	74.1
NeuralSym [2]	44.4	65.3	72.4	78.9
G2Gs [3]	48.9	67.6	72.5	75.5
Transformer [3]	37.9	57.3	62.7	/
Syntax Correction [7]	43.7	60.0	65.2	68.7
Latent model, *l*=1 [8]	44.8	62.6	67.7	71.7
Latent model, *l*=5 [8]	40.5	65.1	72.8	79.4
MSNR	53.0	68.4	72.2	75.6
MSNR+	**63.9**	**83.1**	**87.6**	**91.3**

Table 2 Top-k exact match accuracy in USPTO-full dataset.

Methods	USPTO-50k		USPTO-full	
	Top-1	Top-10	Top-1	Top-10
Retrosim [1]	37.3	74.1	32.8	56.1
Neuralsym [2]	44.4	78.9	35.8	60.8
MSNR	53.0	75.6	37.7	56.7
MSNR+	**63.9**	**91.3**	**51.8**	**79.0**

Retrosynthesis prediction results yielded by each method. Left: the top-k exact match accuracy in the USPTO-50k dataset; Right: the top-k exact match accuracy in the USPTO-full dataset.