A Recurrent Neural Network for molecular structures generation from ^{13}C NMR data

Saúl H. Martinez-Treviño,[1] Gabriel Merino,[1] Victor Uc-Cetina.[2]

[1]Departamento de Física Aplicada, Centro de Investigación y de Estudios Avanzados, Km. 6 Antigua carretera a Progreso Apdo. Postal 73, Cordemex, 97310, Mérida, México

[2]Facultad de Matemáticas, Universidad Autónoma de Yucatán, Av. Industrias no contaminantes, S/N, 97119 Mérida, Yucatán, México.

Structure elucidation of chemical compounds is a very complex and challenging activity that requires some expertise, creativity and well-suited tools. In order to assign the correct molecular structure of a certain compound, NMR is one of the most adopted techniques due to its wide range of structural information.[1] In this way, the exhaustive possibilities within the chemical space are reduced and restricted given the spectroscopic data. With respect to the chemical space exploration, current deep neural networks architectures have been developed in order to generate molecular structures restricted to certain properties. Mainly, in the drug discovery field, there have been several reports[2] of generative models based on neural networks consisting of different arrangements and representations of molecules. Most of the architectures are based on VAE (Variational Autoencoders), GAN (Generative Adversarial Networks), RNN (Recurrent Neural Networks), among others. Given that the search space of the mentioned works allow a wider range of molecules than the spectroscopic restrictions, we want to test the capability of a generative model based solely on spectroscopic data. Thus, the pattern recognition of substructures from the model could help to elucidate the molecular structure. Furthermore, there are no reports about generative models from spectroscopic data. So, we propose a neural network design based on a RNN that generates molecular structures given the NMR data. In this work we present a neural network that consists in a Fully-Connected architecture and a RNN. The input space is the experimental ^{13}C NMR and the output is a molecular structure codified via deepSMILES.[3] We tested the model via 4 main entities: train error, test error, samples prediction accuracy, and functional groups prediction F1-score. Also, we explored the dependence of the proposed model on training size, molecular size, and the experimental environment.

(1) Breton, R.C.; Reynolds, W.F., *Nat. Prod. Rep*., **2013**, *30*, 501.
(2) Brown, N.; Fiscato, M.; Segler, M. H. S.; Vaucher, A. C. *J. Chem. Inf. Model.* **2019**, *59*, 1096.
(3) Noel M. O'Boyle and Andrew Dalke, "DeepSMILES: An Adaptation of SMILES for Use in Machine-Learning of Chemical Structures," in ChemRxiv (2018).