Predicting molecular properties using Recurrent Neural Networks under data scarcity scenarios

Amaia Elizaran Mendarte[1], Gustavo Ariel Schwartz[1]
[1]Centro de Física de Materiales (CSIC-UPV/EHU), Paseo de Manuel Lardizabal, 5, 20018, Donostia-San Sebastián, Gipuzkoa, Spain

aelizaran005@ikasle.ehu.eus

The design of new advanced materials is nowadays broadening its possibilities through Artificial Intelligence [1]. Especially, artificial neural networks are able to capture complex relationships between molecular structures and properties. This advantage significantly reduces the time and cost that we face when designing materials by means of a classical approach.

Many studies have made progress in these advancements throughout the last decade [2], [3], [4]. However, neural networks work well when the purposes they are applied for have plenty of available data, but there are some cases where the available data is scarce or poor-quality data, which affects badly to the performance of the neural networks [5]. Therefore, we specially focus on giving a solution to the problem of data scarcity.

In this work, we use recurrent neural networks (RNN) to predict molecular properties only from the knowledge of the corresponding molecular structures of a scarce database. The SMILES representations of the molecular structures, which are string character sequences, are fed into the algorithm as an input, together with the desired property values. **Figure 1** schematically shows the flow chart of the data processing.

When it comes to tackling data scarcity, we have analyzed different approaches. The focus has been set on similarities among the data and we have created smaller datasets of similar molecules. Then, these datasets have been used to train the network and obtain the desired predictions. The analyzed similarities are of distinct nature. On the one hand, we have considered string similarities of the SMILES encodings. On the other hand, we have computed the similarities of the chemical structures in the vector space (feature space) that is created in the last hidden layer while training. In this way we show how these approaches allow us to improve the performance of the RNN under data scarcity scenarios.

References

[1] Yu-Chen Lo, Stefano E. Rensi, Wen Torng, Russ B. Altman. Machine Learning in chemoinformatics and drug discovery. Drug Discovery Today, 8, pages 1538-1546 (2018). DOI: https://doi.org/10.1016/j.drudis.2018.05.010

[2] Borredon, C., Miccio, L. A., Cerveny, S., Schwartz, G. A. Characterising the glass transition temperature-structure relationship through a recurrent neural network. J. Non-Crystalline Solids: X, 18 (2023). DOI: https://doi.org/10.1016/j.nocx.2023.100185

[3] Walters, W. P. and Barzilay, R. Applications of Deep Learning in Molecule Generation and Molecular Property Prediction. Acc. Chem. Res. 54, 263-279 (2021). DOI: 10.1021/acs.accounts.0c00699

[4] Wenbo Sun et al., Machine learning-assisted molecular design and efficiency prediction for high-performance organic photovoltaic materials. Sci. Adv. 5, eaay4275 (2019). DOI: 10.1126/sciadv.aay4275

[5] Alzubaidi, L., Bai, J., Al-Sabaawi, A. *et al.* A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *J Big Data* 10, 46 (2023). DOI: https://doi.org/10.1186/s40537-023-00727-2

Figures

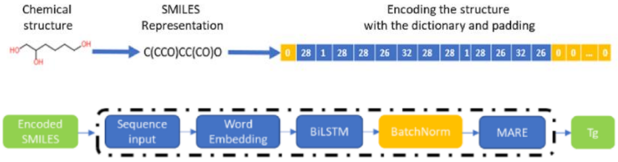

Figure 1. Scheme of encoding of chemical structure and RNN [3].