Similarity-driven framework for data-efficient polymer property prediction

Amaia Elizaran Mendarte[1,2], Gustavo A. Schwartz[1]

[1]*Centro de Física de Materiales (CFM-CSIC/EHU), Paseo Manuel Lardizabal 5, 20018 Dononstia – San Sebastián, Spain*

[2]*Universidad del País Vasco / Euskal Herriko Unibertsitatea (UPV/EHU)*

Abstract

Prediction of polymer properties is crucial in designing new, advanced materials. In recent years, artificial neural networks (ANNs) have become increasingly popular for quantitative structure-property relationship (QSPR), which enables property prediction directly from polymers' structures. However, data scarcity significantly affects the applicability of these models and still remains a major limitation in the field [1]. Here, we present a data-efficient method to tackle data scarcity and enhance performance of an ANN in predicting glass transition temperature (T_g). We extended the similarity principle [2], which states that molecules with similar structures are expected to exhibit similar properties, to develop a similarity-driven framework for data-efficient polymer property prediction [3]. The chemical similarity method incorporates SMILES representations and T_g values to obtain vector representations that are then used to compute the similarity to a target polymer. Once the most similar polymers to the target are selected, a localized regression is applied to their T_g values, as a means to enable predictions with as few as 5 known values. On average the MAPE error obtained among the target polymers outperforms a baseline model. Furthermore, a real improvement is seen for the prediction of those that are outliers in the baseline model. These findings demonstrate that the proposed data-efficient strategy for addressing data scarcity provides accurate predictions of polymers' T_g, outperforming conventional ANN approaches used for QSPR applications.

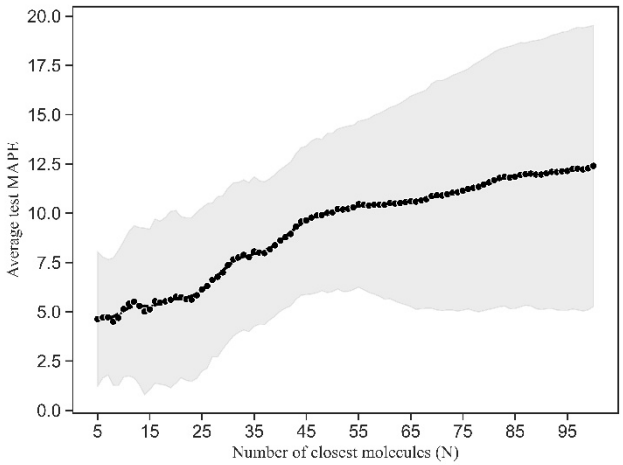

Figure 1: Average Mean Absolute Percentage Error (MAPE) depending on the number of most similar polymers selected for localized regression.

Bibliography :

[1] A.Gangwal et al., *Comput. Biol. Med,* 2024, 179, 108734. https://doi.org/10.1016/j.compbiomed.2024.108734
[2] A. Bender, R.C. Glen, ***Org. Biomol. Chem.***, 2004, 2, 3204-3218. https://doi.org/10.1039/B409813G
[3] Y. Chen, et al., ***J. Mater. Chem. A.***, 2024, 12, 30249-30268. https://doi.org/10.1039/D4TA06452F