MolPILE - large-scale dataset for molecular representation learning

Jakub Adamczyk[1], Jakub Poziemski[2], Franciszek Job[1], Mateusz Król[1], Maciej Makowski[2],

[1]*Faculty of Computer Science, AGH University of Krakow, al. Mickiewicza 30, 30-059, Cracow, Poland.*

[2]*Institute of Biochemistry and Biophysics, Polish Academy of Sciences, ul. Adolfa Pawińskiego 5a, 02-106, Warsaw, Poland*

Abstract Text:

The size, diversity, and quality of pretraining datasets critically determine the generalization ability of foundation models. Despite their growing importance in chemoinformatics, the effectiveness of molecular representation learning has been hindered by limitations in existing small molecule datasets.

To address this gap, we present MolPILE [1]: a large-scale, diverse, and rigorously curated collection of 222 million compounds, constructed from 6 large-scale databases using an automated curation pipeline. We present a comprehensive analysis of current pretraining datasets, highlighting considerable shortcomings for training ML models, and demonstrate how retraining existing models on MolPILE yields improvements in generalization performance. This work provides a standardized resource for model training, addressing the pressing need for an ImageNet-like dataset in molecular chemistry.

Bibliography :

[1] Adamczyk, Jakub, et al. "MolPILE--large-scale, diverse dataset for molecular representation learning." arXiv preprint arXiv:2509.18353 (2025).