LABORATORY FOR CHEMICAL TECHNOLOGY
Technologiepark 125 9052 Gent, Belgium

Coach	**Supervisor(s)**	**Funding**
Maarten Dobbelaere	Kevin Van Geem Maarten Sabbe	FWO

Conformational Data Augmentation for Machine Learning-Based Molecular Property Predictions with Chemical Accuracy

Aim

The aim of this master thesis is to upgrade the accuracy of existing molecular property prediction by including conformational information so that chemical accuracy can be obtained with minimal data.

Justification

Accurately predicting properties of molecules is one of the most critical tasks when designing chemical processes, developing materials, or designing drugs. Machine learning has proven to be an interesting tool to achieve high prediction accuracy with high speed. Typically, property prediction models are either two-dimensional or three-dimensional. Two-dimensional models, also called topology-based models, are described by the molecular graph to which atomic features are added. Three-dimensional or geometry-based models use the molecular geometry of a single conformer. The use of 3D information in property prediction is both a strength and a limitation, because it is time-consuming to calculate a conformer geometry for all datapoints. In existing methods, geometry optimization is avoided by on-the-fly 3D geometry generation such as with the ETKDG method in the cheminformatics toolkit RDKit [1]. This generated conformer is usually not the one with the lowest energy and the inconsistent calculation of geometries leads to a lower prediction accuracy [2]. So-called 4D methods consider molecules as a conformational ensemble. Then, multi-instance learning [3] or self-supervised learning can be used to augment the geometry data, to lower the prediction error and to enable machine learning-based lowest-energy conformer search. Quantum-chemical calculations will be performed as part of the thesis to generate a benchmark database, on which the accuracy of various geometry generation and property prediction methods will be evaluated. An existing 3D property prediction tool will be extended to include a fourth dimension by using generative deep learning algorithms.

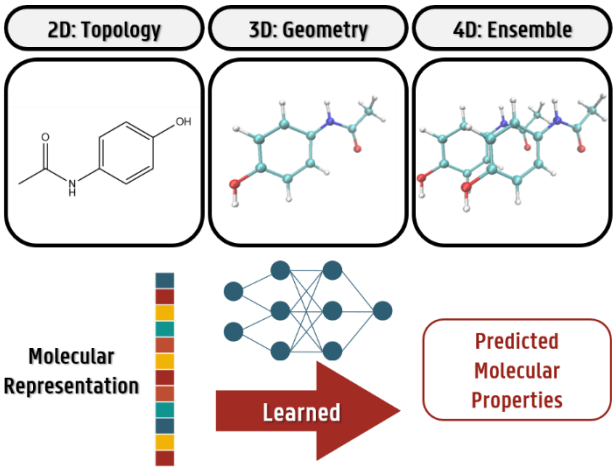

Figure 1: Overview of different dimensions of molecular representations and the workflow for molecular property prediction

Program

- Literature survey on molecular geometry optimization
- Ab initio data generation
- Data curation and database construction
- Machine learning model development and extension of existing models (python)

References

1. Riniker, S. and G.A. Landrum, *Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation.* Journal of Chemical Information and Modeling, 2015. **55**(12): p. 2562-2574.
2. Dobbelaere, M.R., et al., *Learning Molecular Representations for Thermochemistry Prediction of Cyclic Hydrocarbons and Oxygenates.* The Journal of Physical Chemistry A, 2021. **125**(23): p. 5166-5179.
3. Zankov, D.V., et al., *QSAR Modeling Based on Conformation Ensembles Using a Multi-Instance Learning Approach.* Journal of Chemical Information and Modeling, 2021. **61**(10): p. 4913-4923.

GHENT UNIVERSITY
LCT DRIVING CHEMICAL TECHNOLOGY