Data Driven, Explainable Machine Learning Models for Accurate Thermodynamic Predictions

José Ferraz-Caetano[1]

jose.caetano@fc.up.pt · www.jfcaetano.com

Supervisors: Filipe Teixeira[2], M. Natália D. S. Cordeiro[1]

1. LAQV-REQUIMTE – Department of Chemistry and Biochemistry – Faculty of Sciences, University of Porto – Rua do Campo Alegre, S/N, 4169-007 Porto, Portugal
2. CQUM, Centre of Chemistry, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

U.PORTO FACULDADE DE CIÊNCIAS UNIVERSIDADE DO PORTO

LAQV requimte LABORATÓRIO ASSOCIADO PARA A QUÍMICA VERDE

MIT Portugal 2023 Annual Conference

Data Science
In industrial manufacturing

Chemical Industry 4.0
aims for
Sustainable Industrial Process Research
through
Maximize Operations
Data Improves Industrial R&D

Maintenance Calculations
95 % improvement margin

Performance Monitoring
77 % workflow optimization

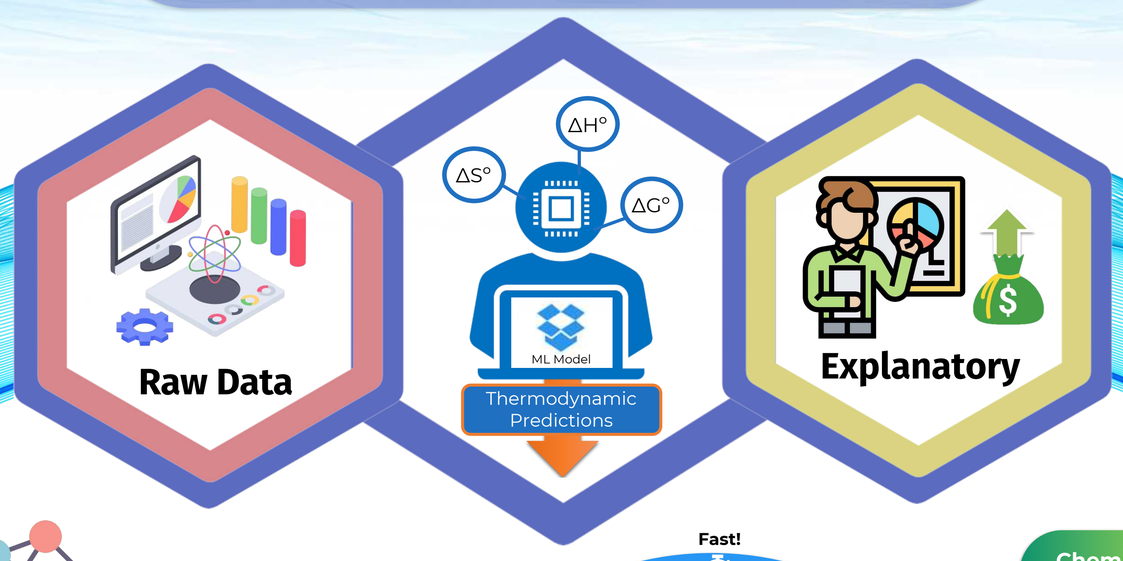
Value for industry data!

Our Goal
A Universal tool to troubleshoot reaction design

Tunable Model for Prediction of Complex Thermodynamic Properties

ΔS° ΔG° ΔH°

We present a Proof of Concept of a Machine Learning Model to predict $\Delta_{vap}H_m^\circ$



The Challenges

Industrial property determination is consistently challenging

due to

High Computational Cost

Seldomly accurate predictions

Narrow chemical application

We overturn this by presenting a **low-cost, explanatory solution** !

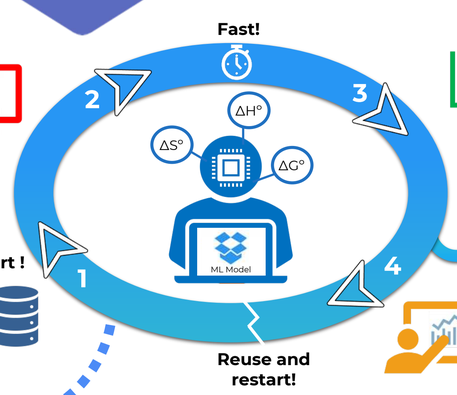
Fast, reliable, replicable!

How it works

- Database comprises 1422 different molecules
- Over 20 organic/inorganic functional groups
- Using 105 open-source molecular descriptors

In House Software

2 Machine Learning — Supervised Ensemble Algorithms



Chemical Features 3

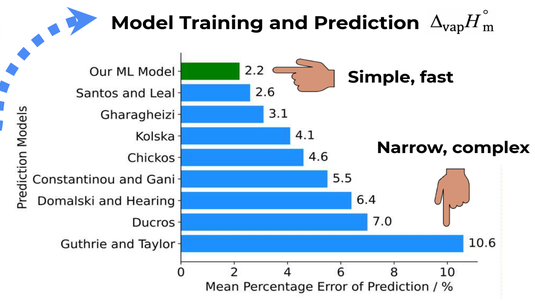
Prediction! $\Delta_{vap}H_m^\circ$

Chemical Explanation 4

MOL file — OpenBabel — SMILES — RdKit — Descriptor — 1 Chemical Database

We built an in-house framework for descriptor generation by amassing open-source experimental and topological data.

Model Training and Prediction $\Delta_{vap}H_m^\circ$



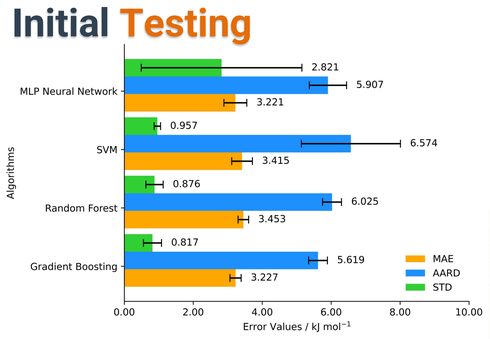
ML model predicted the standard vaporization enthalpies of a external database **with a better MPE than previous benchmark attempts !**

The model was successfully validated with $\Delta_{vap}H_m^\circ$ calculation for an external database.

We can use explanatory data not only to predict properties, but also... **predict structures** !

Our Workflow

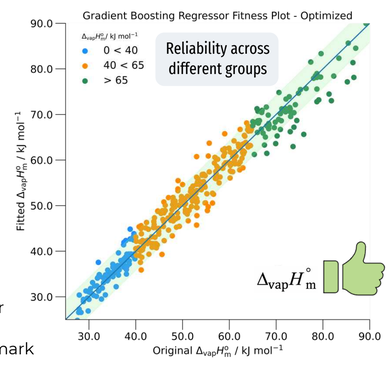
Initial Testing



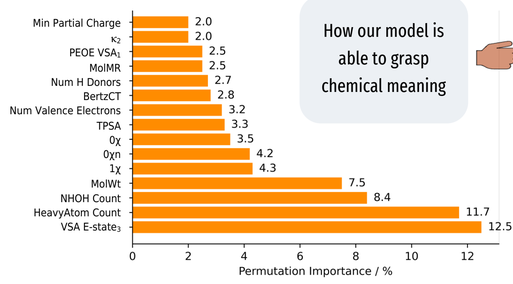
MAE = 3.158 kJ mol^{-1} Well **below** the 4 kJ mol^{-1} barrier

STD: 0.445 kJ mol^{-1} **Higher** than 3.1 kJ mol^{-1} benchmark

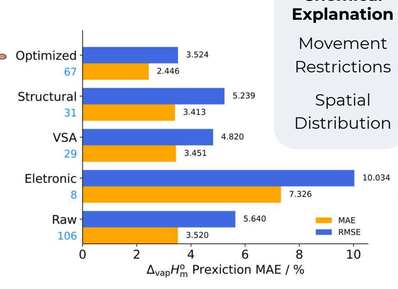
Model Optimization



Feature Importance



How our model is able to grasp chemical meaning



Chemical Explanation
Movement Restrictions
Spatial Distribution

Molecules with a **smaller surface area** of **low electronic and topological environment**, resulting in and **smaller van der Waals forces**, seem to provide better predictions!

MAE = 2.446 kJ mol^{-1}

Well below the 4 kJ mol^{-1} barrier

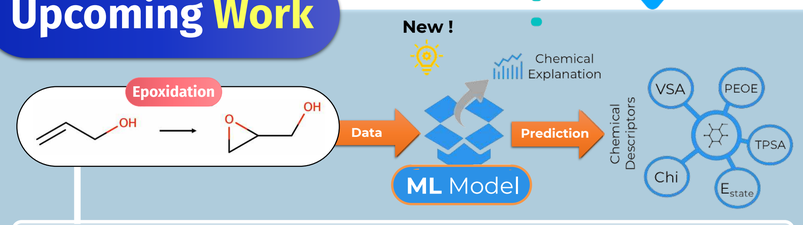
Using **Supervised Machine Learning Algorithms** (Random Forest and Gradient Boosting), our **model outperformed benchmark studies**, naming key **chemical properties behind each prediction**.

Major Method Outputs

Our models are widely more accurate than previous ML published methodologies for thermochemical property prediction.

This methodology yields a replicable model framework which can be expanded for the prediction of other thermodynamic properties.

Upcoming Work



Our MIT Portugal PhD Project will use this model as a key component for chemical reaction prediction !

Funded by:

fct Fundação para a Ciência e a Tecnologia

MIT Portugal

Graphics and elements used with *Venngage*, *Slidesgo* and *Flaticon* commercial licenses. | Pilot 'Proof of Concept' Study journal publishing is forthcoming. | **Economic data**: *Science of the Total Environment*, 2017 (598) 931–936.

Acknowledgements: The authors thank the Fundação para a Ciência e Tecnologia (FCT/MCTES) support to LAQV-REQUIMTE (UIDP/50006/2020). José Ferraz-Caetano's PhD Fellowship is supported by a doctoral Grant (SFRH/BD/151159/2021) financed by the FCT, with funds from the Portuguese State and EU Budget, through the Social European Fund and Programa Por_Centro, under the MIT Portugal Program.

Under the Doctoral Grant SFRH/BD/151159/2021 | **Data Science**