Stiv Llenga | 08.02.2024

Deep Learning – the Death of Physics-Based Kernel Representations?

Physics-based representations for machine learning properties of chemical reactions[1]

- Reaction representations are constructed using conventional QML representations, accounting for the nuclear charge and coordinates of the species involved in the reaction.
- A new reaction representation (B^2R^2) is created by investigating the traditional representation ingredients relevant to high-performing reactions.

Method:

- Reactant-only, product-only, or combined (addition or difference) representations of both can be used to represent a chemical reaction.

$$X_{reactants(products)} = \sum_{i(j)=1}^{N(M)} X^{i(j)} \qquad X_{\text{difference}} \approx X_{TS} = \sum_{\text{j=1}}^{\text{M}} X^{\text{products}} - \sum_{\text{i=1}}^{\text{N}} X^{\text{reactants}}$$

- Analysis of diverse datasets showed that combined representations perform better.
- The analysis of the key factors that contribute to the robustness of SLATM_d, namely its meaningful difference, emphasis on two-body interactions, and separation of relevant two-body feature bags, is used to engineer the B^2R^2 representation.
- B^2R^2 is based on the notion of differences in pairwise interactions between reactants and products;
- Three bagging strategies are used by B^2R^2: canonical (pairwise bags), linear (elemental bags), and constant-sized (no bags).

Acceptable accuracy is achieved when the B^2R^2 representation is used as input for KRR models when predicting reaction properties.

[1] *Mach. Learn.: Sci. Technol.* **2022**, *3*, 045005.

Critical comments[2]

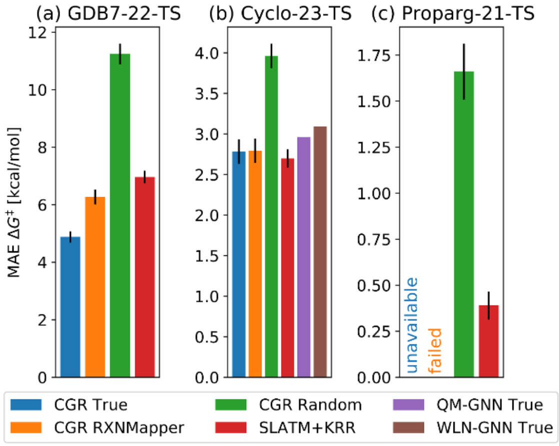
- Model performance: For quantitative activation energy analysis, a model with high test MAE (*ca.* 10 kcal/mol) is useless. Using model baselines to compare performance is essential.
- Model input: For reaction energies, the output is computed at no extra cost because the model uses DFT-optimised geometries. Methods cheaper than DFT must be used to incorporate 3D information (only if really needed).
- Data splitting: Training models on random splitting can deduce irrelevant patterns, hindering generalization. Using different splits (including chemically relevant ones) can help determine applicability and compare this model to others.

Splitting	SLATM_d(2)	DMPNN
Random	10.01±0.16	4.11±0.07
Scaffold	13.32±0.73	19.08±0.94

MAEs (kcal/mol) are reported for 10,165 training points of GDB7-20-TS dataset.

Reply to critical comments[3]

- Model input and performance: Deep NN models use atom-mapped reaction SMILES, and the mapping method (hand-made or algorithmically evaluated) greatly affects model performance. Both "model types" have advantages and disadvantages, and different datasets present different challenges.
- Data splitting: All dataset reactions appear in the test set in one of CV folds. Community consensus on best practices is lacking.



[2] *Mach. Learn.: Sci. Technol.* **2023**, *4*, 048001.

[3] Image reproduced from *Mach. Learn.: Sci. Technol.* **2023**, *4*, 048002 under CC BY license (OA).