Advancing *ab initio* materials modelling using symbolic regression and large language models

Amit Chaudhari *[a] and Andrew Logsdail [a]
[a] Cardiff Catalysis Institute, School of Chemistry, Cardiff University, Cardiff CF10 1AT, UK
* ChaudhariA@Cardiff.ac.uk

CECAM Machine Learning for Materials Discovery workshop, May 2025, Espoo

Accelerating Materials Modelling Using "*Deorbitalised*" Meta-GGA Density Functional Theory

- "*Deorbitalised*" meta-GGA density functional theory can accelerate the simulation of molecules and periodic solids, **enabling more expansive *ab initio* materials modelling, *e.g.*, molecular dynamics [1] and QM/QM embedding [2].**
- The accuracy and transferability of deorbitalisation schemes depends on the accuracy of the underlying **kinetic energy density functional (KEDF),** which is a **semi-local approximation for the non-local kinetic energy density (τ)** using at least the electron density (ρ) and the density gradient norm ($\nabla\rho$).
- **τ is highly non-linear and therefore difficult to predict using common supervised ML methods** without introducing a Laplacian-dependence which can introduce numerical instability; thus **we need more accurate methods for regression**.

Integrating Symbolic Regression and Large Language Models to Construct a Kinetic Energy Density Functional

2,240,703 Integration Grid Point Evaluations: 50 Molecules and Periodic Solids

Training Set: Stratified Sample by $\log_{10}(\tau^{DFT})$

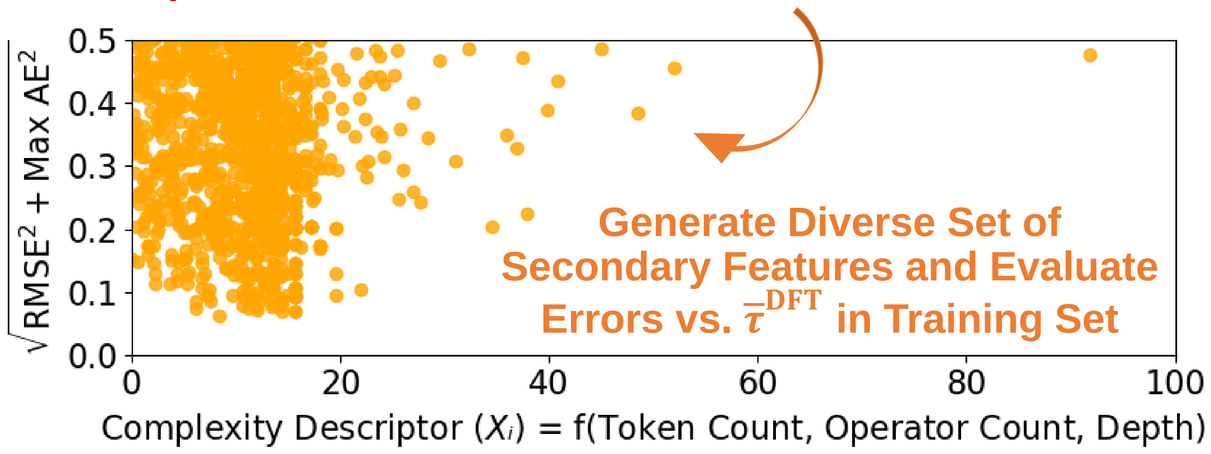
$\bar{\tau}^{DFT}$	$\bar{\rho}$	$\overline{\nabla\rho}$	$\bar{\tau}^{TF}$	$\bar{\tau}^{VW}$
…	…	…	…	…

Target — 4x Primary Features

Secondary Feature Generation (X)

Exhaustive (e.g., SISSO [1]) — ❌ Depth ≤ 3

Random Subset (Python) — ✅ Depths 1-10



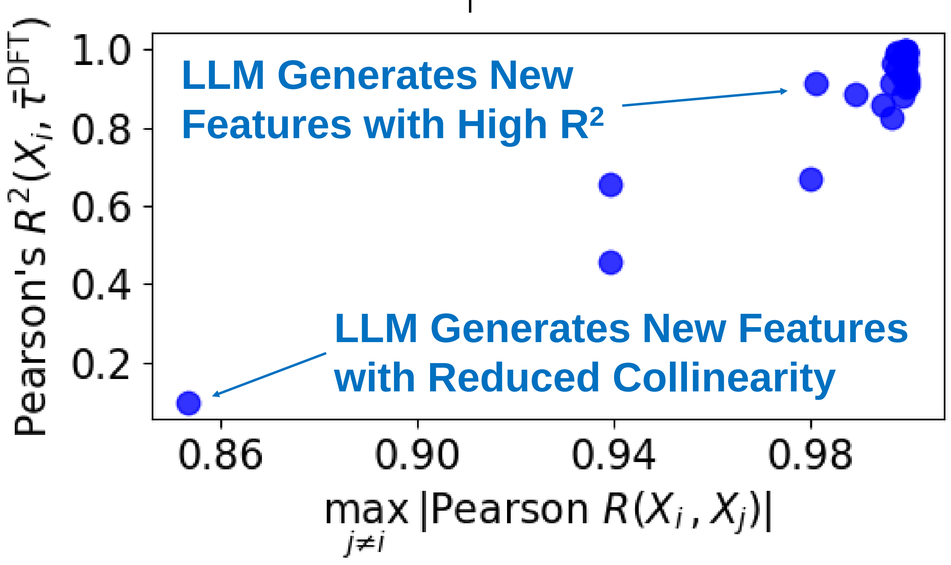
Seed LLM With X_i (i=1-8055), Errors and R^2

Prompt: Use your expertise in symbolic regression and natural language processing to generate new symbolic expressions, that involve the same primary variables and different algebraic operators including addition, subtraction, multiplication, division, exponentials, powers, logarithms and trigonometric functions, with different expression "depths" …

Natural Language Processing Task

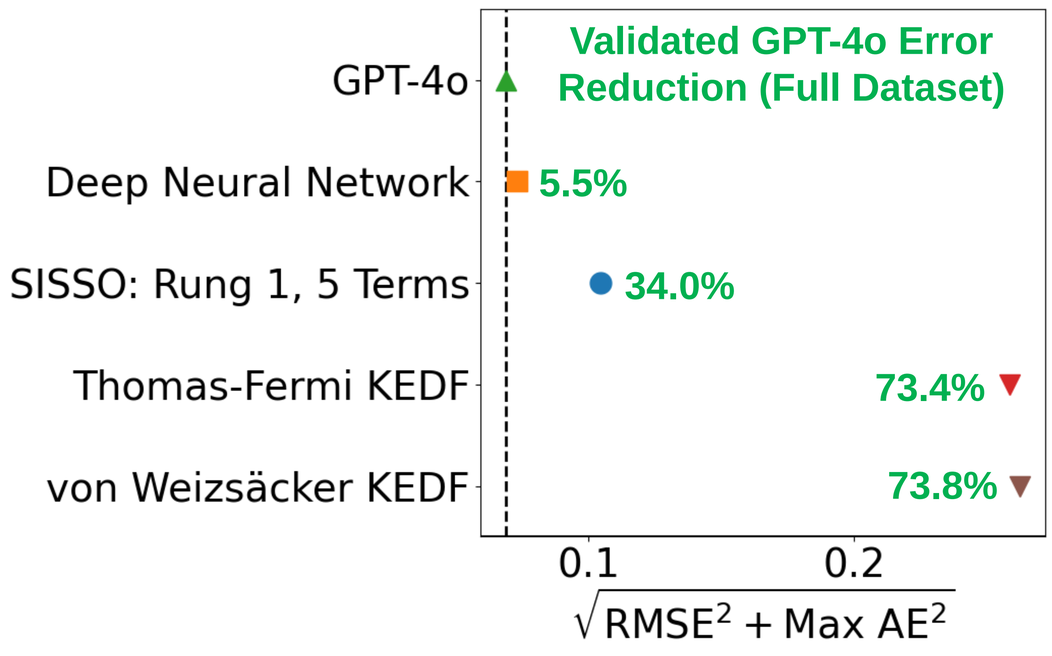
OpenAI GPT-4o Large Language Model

Generate New Candidate Secondary Features X_{i-j}



Linear Regression to Predict $\bar{\tau}^{DFT}$

Test Every Combination of 3 Secondary Features in X_{i-j} → Optimised KEDF



$$\bar{\tau}^{DFT} = c_0 + c_1 T_1 + c_2 T_2 + c_3 T_3$$

$$T_1 = \frac{\left(\bar{\tau}^{TF} - \bar{\tau}^{VW}\right)^2}{1 + \bar{\rho} + \overline{\nabla\rho}} \quad T_2 = \frac{\cos\left(\overline{\nabla\rho} \times \bar{\tau}^{TF}\right)}{\sqrt{1 + \bar{\tau}^{VW}}} \quad T_3 = \frac{\bar{\tau}^{VW}}{1 + \log_{1p}\left(\bar{\rho} \times \bar{\tau}^{TF}\right)}$$

[1] D. Mejia-Rodriguez and S. B. Trickey. *Deorbitalized meta-GGA exchange-correlation functionals in solids*. 2018. https://doi.org/10.1103/PhysRevB.98.115161

[2] W. Mi, X. Shao, A. Genova, *et al. eQE 2.0: Subsystem DFT beyond GGA functionals*. 2021. https://doi.org/10.1016/j.cpc.2021.108122

[3] R. Ouyang, S. Curtarolo, E. Ahmetcik, *et al. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates*. 2018. https://doi.org/10.1103/PhysRevMaterials.2.083802

References and Acknowledgements

Cardiff University / Prifysgol Caerdydd; International Centre for Advanced Materials; GW4; Supercomputing Wales / Uwchgyfrifiadura Cymru; ARCHER2; JM; bp; EPSRC Pioneering research and skills; HEC MCC; FHI-aims The ab initio materials simulation package