Transferable diversity – a data-driven representation of chemical space.

Stephen G. Dale[1], Tim Gould[2], Bun Chan[3],
Stefan Vuckovic[4]
[1] Institute for Functional Intelligent Materials,
National University of Singapore,
Block S9, Level 9, 4 Science Drive 2, Singapore 117544
[2] Queensland Micro- and Nanotechnology Centre,
Griffith University,
Nathan, Qld 4111, Australia
[3]Graduate School of Engineering,
Nagasaki University,
Bunkyo 1-14, Nagasaki 852-8521, Japan
[4]Department of Chemistry,
University of Fribourg,
Chem. du Musée 9, 1700 Fribourg, Switzerland

sdale@nus.edu.sg

Transferability, especially in the context of model generalization, is a paradigm of all scientific disciplines. However, the rapid advancement of machine learned model development threatens this paradigm, as it can be difficult to understand how transferability is embedded (or missed) in complex models developed using large training data sets. Two related open problems are how to identify, without relying on human intuition, what makes training data transferable; and how to embed transferability into training data. To solve both problems for *ab initio* chemical modelling, an indispensable tool in everyday chemistry research, we introduce a *transferability assessment tool* (TAT) and demonstrate it on a controllable data-driven model for developing density functional approximations (DFAs). We reveal that human intuition in the curation of training data introduces chemical biases that can hamper the transferability of data-driven DFAs. We use our TAT to motivate transferability principles; one of which introduces the key concept of transferable diversity. Finally, we propose data curation strategies for general-purpose machine learning models in chemistry that identify and embed the transferability principles.[1]

References

[1] Stefan Vuckovic, Tim Gould, Bun Chang, Stephen Dale, chemrxiv, (2023).