# TRAITS: TRusted AI Toolkit for Science

Ashley S. Dale*, Daniel Persaud*, Kangming Li*,
Yao Fehlis^, Brian DeCost‡, Jason Hattrick-Simpers*

*Department of Materials Science and Engineering, University of Toronto; ^Artificial, Inc.,
‡National Institute of Standards and Technology

Materials Science & Engineering
UNIVERSITY OF TORONTO

Schmidt Sciences

## Case Study: Are these models overconfident? Should we trust them?

*Overconfidence*: *low uncertainty for incorrectly predicted values.*

**Task**: Predict the band gap energies without training on datasets that contain all atomic elements

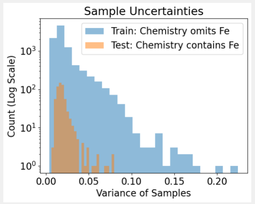
**Method**: Train an ALIGNN model [1] on chemistries from the without element X. Test the ALIGNN model using chemistries containing element X. Let element $X \in [Fe, F]$ [2]. Confidence scores were estimated using Hessian-Weight Averaged Gaussian method[3].

## Results for Fe-Omission Study







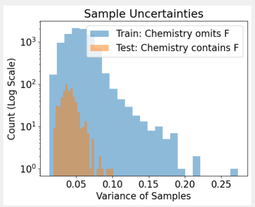

The model is over-confident: well-predicted and poorly-predicted samples have similar uncertainty. This can be probed by perturbing model weights and observing changes in the loss landscape.

## Results for F-Omission Study





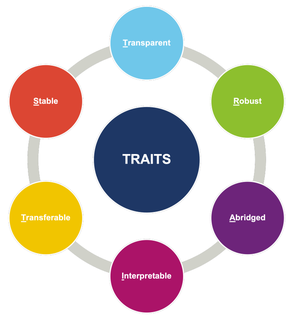



This model is also over-confident and untrustworthy. The analysis does not provide insight into why.

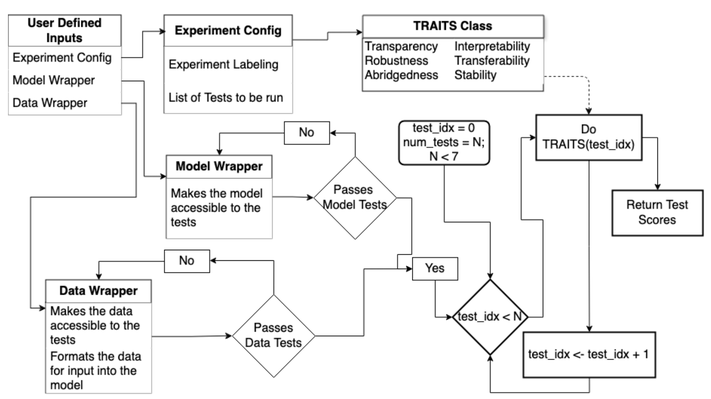
## TRAITS Overview

The **TRusted AI Toolkit for Science** is a model agnostic and task agnostic Python package inspired by various Trustworthy AI Frameworks [4]. Common methods for evaluating the trustworthiness of a model – such as the case study presented here – may fail to provide insight into the source of the untrustworthiness. TRAITS allows users to answer questions about their model such as




- **Transparent**: Is the model, data, and methodology transparent so that it is possible for a third party to reproduce results?
- **Robust**: Is the model robust to noise in the data, so that if the data quality is degraded, model performance is unchanged?
- **Abridged**: How badly is the model over-parameterized? Can parameters be removed without changing model performance?
- **Interpretable**: Given that simpler models are inherently more interpretable than more complex models, how complex is the model?
- **Transferable**: Given out-of-distribution data, how well does the model generalize?
- **Stable**: How robust is the feature representation learned by the model to perturbation?

## TRAITS Workflow

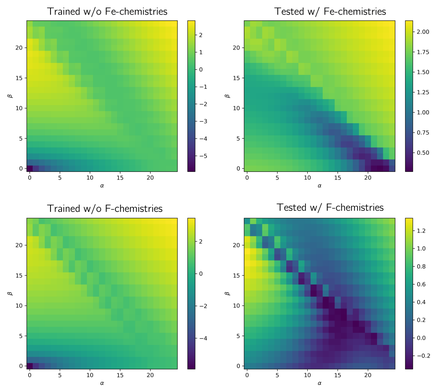



## Search for New Trustworthiness Metrics

Loss landscapes $L(\theta, \alpha i, \beta j)$ along eigenvectors $\alpha, \beta$ are rigorous and computationally tractable when using the *Hessian Vector Product* [5]:

$$\nabla_\theta \left[ (\nabla_\theta L)^T \right] = (\nabla_\theta \nabla_\theta L)\, v + (\nabla_\theta L)^T \nabla_\theta v = H_\theta v$$

Eigenvalues $\kappa^{\alpha,\beta}$ of $H_\theta$ represent principal curvatures of the loss landscape. Curvature is associated with model robustness, complexity, and stability [6].




## Outlook and Next Steps

- Release of TRAITSv1.0 on GitHub and PyPi
- Validation of newly proposed metrics
- Benchmarking SOTA models with TRAITS in addition to performance metrics

## Summary

- Overconfident and untrustworthy AI/ML models are problematic
- This is demonstrated using a model trained to predict bandgap energies for new chemistries
- TRAITS standardizes a flexible set of trustworthiness tests
- Loss landscapes are potentially useful for characterizing trustworthiness

## References

[1] K. Choudhary and B. DeCost, "Atomistic line graph neural network for improved materials property predictions," *npj Computational Materials*, vol. 7, no. 1, p. 185, 2021.

[2] K. Choudhary, "Jarvis-dft 3d dataset," 7 2018.

[3] H. Ravishankar, R. Patil, D. Anand, V. Singhal, U. Agrawal, R. Venkataramani, and P. Sudhakar, "Stochastic weight perturbations along the hessian: A plug-and-play method to compute uncertainty," in *Uncertainty for Safe Utilization of Machine Learning in Medical Imaging* (C. H. Sudre, C. F. Baumgartner, A. Dalca, C. Qin, R. Tanno, K. Van Leemput, and W. M. Wells III, eds.), (Cham), pp. 80–88, Springer Nature Switzerland, 2022.

[4] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou, "Trustworthy ai: From principles to practices," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–46, 2023.

[5] L. Böttcher and G. Wheeler, "Visualizing high-dimensional loss landscapes with hessian directions," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2024, no. 2, p. 023401, 2024.

[6] E. Lau, Z. Furman, G. Wang, D. Murfet, and S. Wei, "The local learning coefficient: A singularity-aware complexity measure," 2024.

GitHub.com/daleas0120/TRAITS

This project is supported by the University of Toronto's Eric and Wendy Schmidt AI in Science Postdoctoral Fellowship, a program of Schmidt Sciences.

Ashley.Dale@utoronto.ca