Supplemental Information: Reproducibility in Materials Informatics: Lessons from 'A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials'

Daniel Persaud,[1,2] Logan Ward,[3] and Jason Hattrick-Simpers[1,2]

[1]*Department of Materials Science and Engineering, University of Toronto, Canada*

[2]*Acceleration Consortium, University of Toronto, Canada*

[3]*Data Science and Learning Division, Argonne National Laboratory, USA*

In this document, we provide the results of the investigation of the pseudo-random seed sensitivity test with a Random Forest (RF) regression model and an XGBoost (XGB) regression model. Any reference to files or scripts in github are provided in square brackets.

A. Validating Models

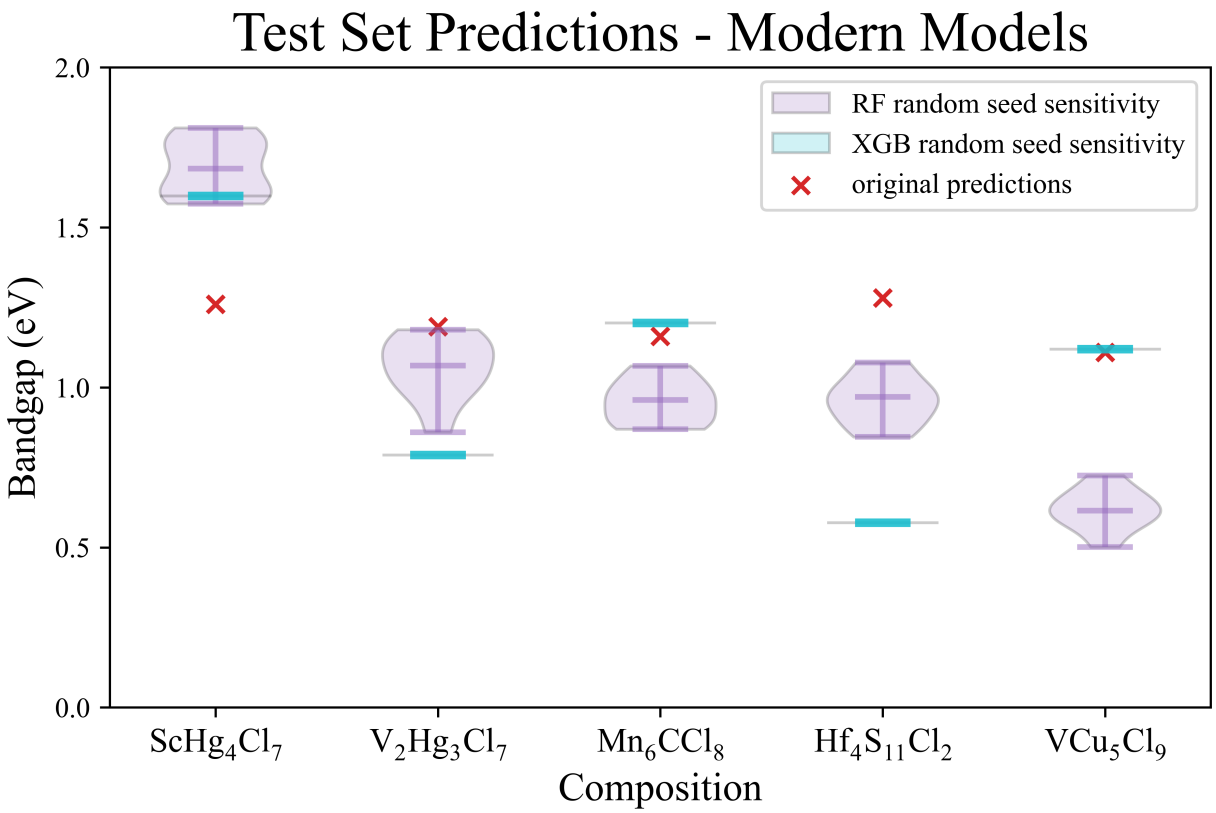
To validate that the model models perform comparably to the hierarchical model, we performed 10-fold random cross validation using a RF model with default parameters (setting the pseudo-random seed to 0), as well as a XGB model with default parameters (setting the pseudo-random seed to 0). All models, including the original hierarchical model are trained with the ICSD entries in OQMD [bandgap.data], for which descriptors are generated in the same way [make-features.in]. The mean RMSE across 10 random folds for all provided in table I. The mean RMSE being within 0.01eV demonstrates the models in distribution performance is comparable.

TABLE I. Mean of the RMSE of each of the 10 folds in a 10-fold random cross validation of the ICSD entries in OQMD

Model	Mean RMSE (eV)
Hierarchical Model	0.686
Random Forest	0.677
XGBoost	0.696

B. Modern Model Pseudo-Random Seed Sensitivity

Once it has been validated that RF and XGB models with default parameters perform similarly to the hierarchical model, all models are initialized with 10 different pseudo-random seeds. All models are trained with the entire training set (OQMD 1.0) and used to predict the test set, just as described in the orignal work. Supplementary figure 1 demonstrates that the RF model is not very sensitive to pseudo-random seed and XGB is not sensitive.



Supplementary Figure 1. The original predictions (red x's) compared to the predictions from the Random Forest pseudo-random seed sensitivity (purple violins) and the predictions from the XGBoost pseudo-random seed sensitivity (cyan violins).