SUPPLEMENTARY INFORMATION FOR ASSESSING DATA-DRIVEN PREDICTIONS OF BAND GAP AND ELECTRICAL CONDUCTIVITY FOR TRANSPARENT CONDUCTING MATERIALS

Uncertainty quantification

Uncertainty quantification is critical for ML models in materials discovery, as experimental validation is resource-intensive. It is essential to model uncertainties in predictions to improve reliability and guide experimental efforts effectively. Uncertainty is typically categorized into two types:

- **Aleatoric uncertainty:** intrinsic noise in the observations, reducible only by improving data quality. It can be homoscedastic (constant variance) or heteroscedastic (variance dependent on specific inputs), with heteroscedastic uncertainty being common in materials science due to varying measurement conditions and sample qualities.
- **Epistemic uncertainty:** Model uncertainty due to insufficient data. It is reducible by incorporating more data in the training process.

Deep learning models do not naturally capture uncertainties, often yielding overconfident predictions. Aleatoric uncertainty in neural-network models can be captured by predicting the parameters of a heteroscedastic Gaussian distribution from the last layer, modeling both the predictive mean $f_\theta(\mathbf{x}_i)$ and variance $\hat{\sigma}^2_{a,\theta}(\mathbf{x}_i)$ using a *Robust* loss function [1; 2]:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \frac{|y_i - f_\theta(\boldsymbol{x}_i)|}{2\hat{\sigma}^2_{a,\theta}(\mathbf{x}_i)} + \frac{1}{2}\log(\hat{\sigma}^2_{a,\theta}(\mathbf{x}_i)). \tag{1}$$

For epistemic contribution to the uncertainty, deep ensembles [3] are used, where the variance across predictions from multiple neural networks approximates the bayesian predictive distribution. The final uncertainty combines aleatoric and epistemic components:

$$\hat{\sigma}^2(\boldsymbol{x}_i) = \hat{\sigma}^2_e(\boldsymbol{x}_i) + \frac{1}{M}\sum_{m=1}^{M} \hat{\sigma}^2_{a,\theta_m}(\boldsymbol{x}_i)\,, \tag{2}$$

where $\hat{\sigma}^2_{a,\theta_m}(\boldsymbol{x}_i)$ denotes the contribution to aleatoric uncertainty produced by the m-th model in the ensemble, while $\hat{\sigma}^2_e(\boldsymbol{x}_i)$ denotes the contribution to the epistemic uncertainty, obtained by computing the variance over predictions from all the models in the deep ensemble. Unlike fully Bayesian methods like Bayesian Neural Networks (BNNs) [4], deep ensembles approximate the Bayesian posterior by training multiple neural networks independently with different random initializations and data shuffling, providing a scalable and practical approximation to bayesian inference.

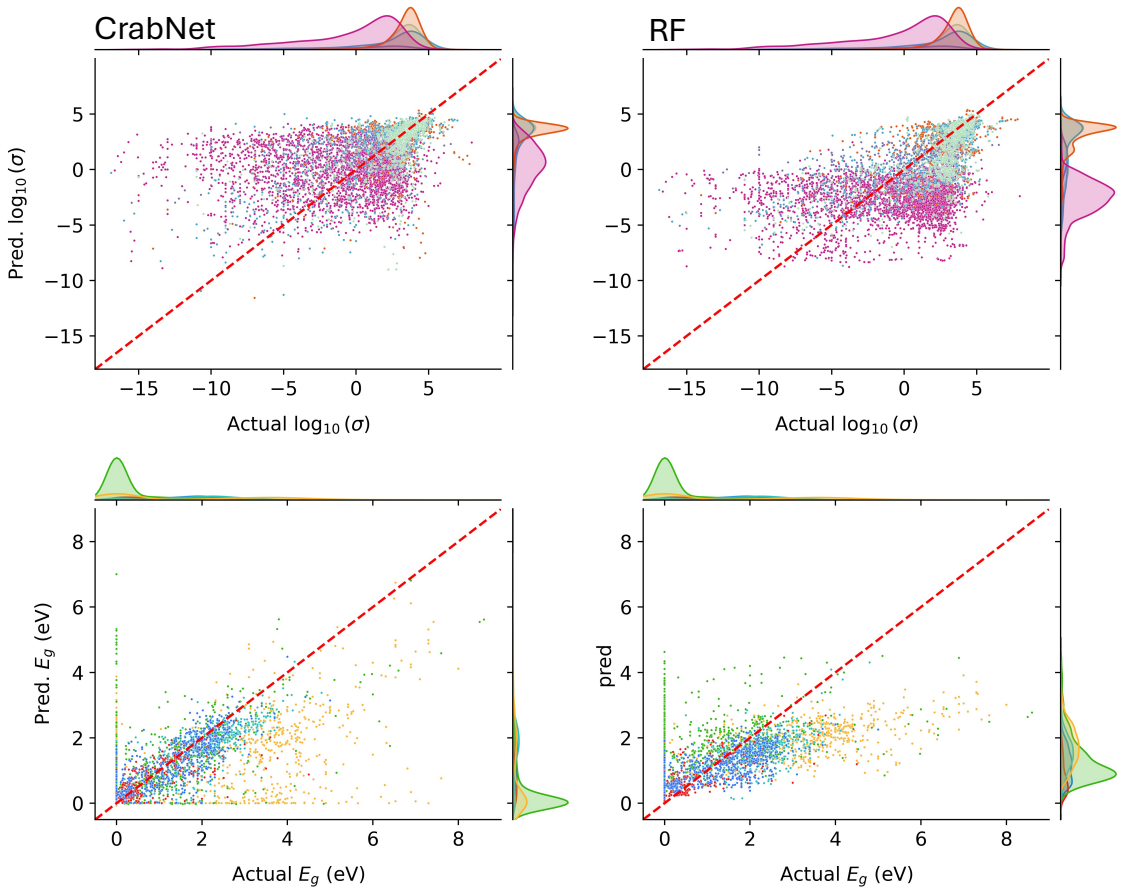

Figure 1: Parity plots of ML cluster predictions under the LOCO-CV evaluation scheme.

LOCO-CV material clusters analysis

In this section, we present a more detailed analysis of material clusters generated using the LOCO-CV [5] evaluation method, as described in Sections of the main thesis. In Figure 1 we show parity plots related to LOCO-CV evaluation scheme, colored according to different material clusters encountered in both conductivity, and band gap datasets. In Figure 2 we report the top-5 element prevalence for each chemical cluster. In general, the presence of diverse, predominant elements in each cluster indicates that the clustering algorithm has successfully grouped the chemical formulas based on their composition. Moreover, the diversity of material groups suggests that the clusters effectively represent distinct regions of the chemical space, potentially capturing different types of compounds or materials.

Conductivity database clusters In the case of conductivity database, **Cluster 0** (Se-Cu-Bi-Sn-Pb) consists of selenium containing compounds as selenides and selenide oxide or selenide halides, while **Cluster 1** (O-Sr-Cu-La-Mn) contains oxides including sulphates and phosphates. **Cluster 2** (Sb-Si-Ge-Ni-Co) is characterized by intermetallic compounds, including borides, carbides, and nitrides, with a significant presence ($\frac{1}{4}$ of entries) of antimony containing compounds. **Cluster 3** (S-Sn-Cu-Bi-Ni) also contains intermetallic compounds along with sulphides, borides, carbides and halide compounds, while approximately $\frac{1}{3}$ of the entries in **Cluster 4** (Te-Fe-Al-Pd-As) consists of materials containing tellurium, with the rest consisting mostly of other intermetallic compounds and some oxide containing materials.

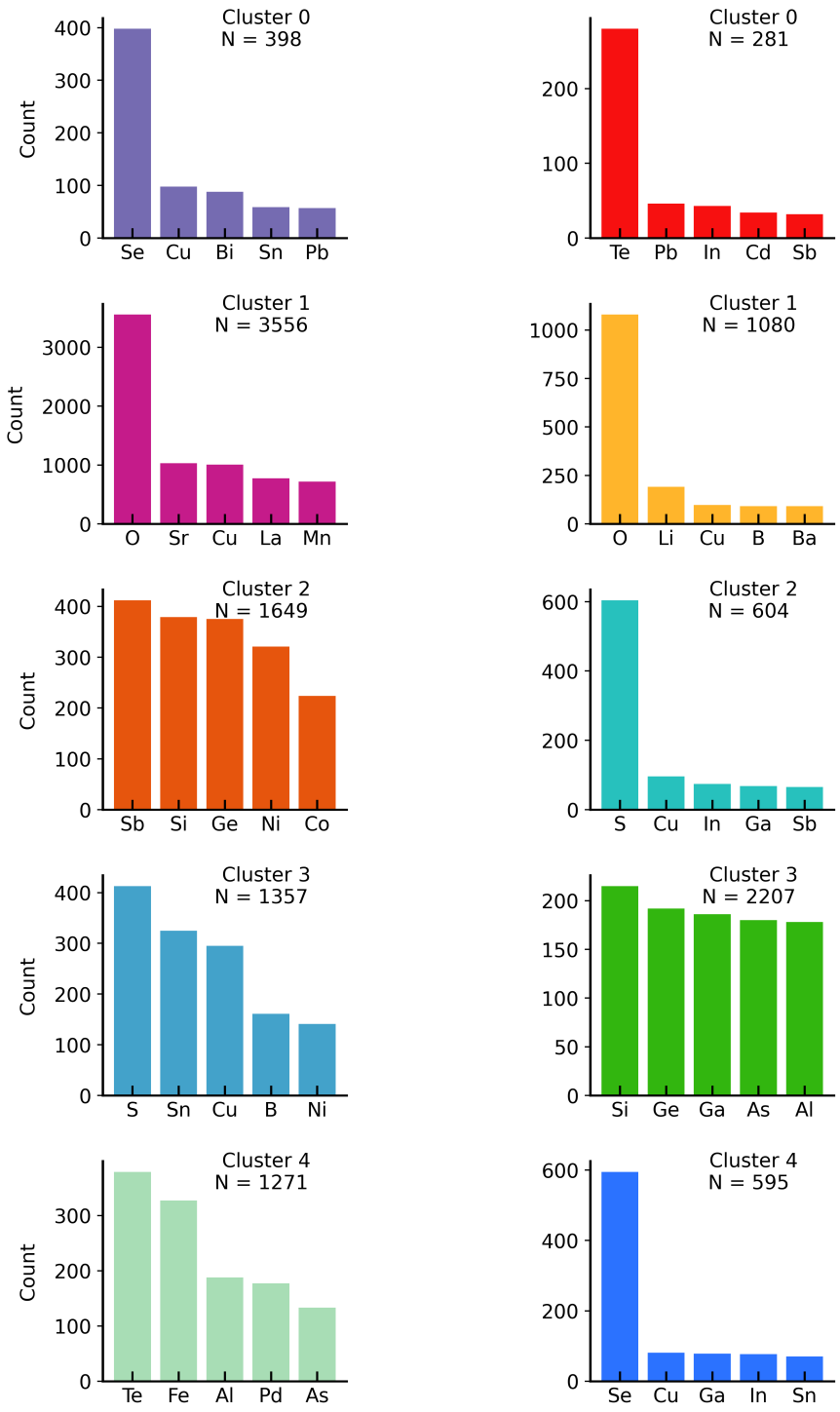

Figure 2: Top-5 element prevalence of LOCO-CV material clusters both for conductivity (**left**) and band gap (**right**) datasets.

Band gap database clusters For the band gap database, **Cluster 0** (Te-Pb-In-Cd-Sb) consists mainly of tellurides and lead-based compositions, while **Cluster 1** (O-Li-Cu-B-Ba) represents oxide containing compounds including sulphates and phosphates. **Cluster 2** (S-Cu-In-Ga-Sb) consists of sulphide materials, including sulphide halides. **Cluster 3**

(Si-Ge-Ga-As-Al) represent intermetallics, including silicides, phosphides, carbides, borides, nitrides, while **Cluster 4** (Se-Cu-Ga-In-Sn) consists of selenide materials including selenide halides.

References

[1] Janosh Riebesell. Probabilistic Data-Driven Discovery of Thermoelectric Materials. *MPhil thesis, University of Cambridge*, 2019. URL `https://github.com/janosh/thermo`.

[2] Rhys E. A. Goodall and Alpha A. Lee. Predicting materials properties without crystal structure: deep representation learning from stoichiometry. *Nature Communications*, 11(1):6280, Dec 2020. ISSN 2041-1723. doi:10.1038/s41467-020-19964-7. URL `https://doi.org/10.1038/s41467-020-19964-7`.

[3] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles, 2016. URL `https://arxiv.org/abs/1612.01474`.

[4] Julyan Arbel, Konstantinos Pitas, Mariia Vladimirova, and Vincent Fortuin. A primer on bayesian neural networks: Review and debates, 2023. URL `https://arxiv.org/abs/2309.16314`.

[5] Bryce Meredig, Erin Antono, Carena Church, Maxwell Hutchinson, Julia Ling, Sean Paradiso, Ben Blaiszik, Ian Foster, Brenna Gibbons, Jason Hattrick-Simpers, Apurva Mehta, and Logan Ward. Can machine learning identify the next high-temperature superconductor? examining extrapolation performance for materials discovery. *Mol. Syst. Des. Eng.*, 3:819–825, 2018. doi:10.1039/C8ME00012C. URL `http://dx.doi.org/10.1039/C8ME00012C`.