# ET-AL: Entropy-Targeted Active Learning for Bias Mitigation in Materials Data

**Preprint** · November 2022

**4 authors:**

**Hengrui Zhang**
Northwestern University
**9** PUBLICATIONS **20** CITATIONS

**Wei Chen**
Siemens
**58** PUBLICATIONS **714** CITATIONS

**James M Rondinelli**
Northwestern University
**319** PUBLICATIONS **15,400** CITATIONS

**Wei Chen**
Northwestern University
**538** PUBLICATIONS **22,003** CITATIONS

**Some of the authors of this publication are also working on these related projects:**

Project — Active Expansion Sampling (AES) View project

Project — Design Space Exploration View project

# ET-AL: Entropy-Targeted Active Learning for Bias Mitigation in Materials Data

**Hengrui Zhang[1], Wei (Wayne) Chen[1], James M. Rondinelli[2,\*], and Wei Chen[1,\*,†]**

[1]Department of Mechanical Engineering, Northwestern University, Evanston, IL 60208, USA
[2]Department of Materials Science and Engineering, Northwestern University, Evanston, IL 60208, USA

**SUMMARY**

Growing materials data and data-driven informatics drastically promote the discovery and design of materials. While there are significant advancements in data-driven models, the quality of data resources is less studied despite its huge impact on model performance. In this work, we focus on data bias arising from uneven coverage of materials families in existing knowledge. Observing different diversities among crystal systems in common materials databases, we propose an information entropy-based metric for measuring this bias. To mitigate the bias, we develop an entropy-targeted active learning (ET-AL) framework, which guides the acquisition of new data to improve the diversity of underrepresented crystal systems. We demonstrate the capability of ET-AL for bias mitigation and the resulting improvement in downstream machine learning models. This approach is broadly applicable to data-driven materials discovery, including autonomous data acquisition and dataset trimming to reduce bias, as well as data-driven informatics in other scientific domains.

**Keywords:** data-driven informatics, materials data, data bias, information entropy, active learning

---

[*] Correspondence: jrondinelli@northwestern.edu (J.M.R.), weichen@northwestern.edu (W.C.)
[†] Lead contact.

## INTRODUCTION

Data-driven autonomous materials design has recently emerged as a new paradigm for materials discovery[1-3]. With large materials data and powerful informatics tools, this paradigm significantly accelerates the understanding of physical and chemical mechanisms in materials science[4-6], accurate prediction of materials structures and properties[7-10], as well as the design of materials with desired properties[11-14]. While the informatics tools, such as machine learning (ML) and design optimization models, hold a conspicuous position in these works, the data resources are as important[15,16]. The performances that the models can attain highly depend on the quality of data they are built upon. Data veracity entails a description of where and how data were collected, but less frequently is why (or why not) using certain data clearly articulated.

Following the Materials Genome Initiative[17], multiple materials data resources have emerged. The Materials Project[18], Open Quantum Materials Database (OQMD)[19,20], and the Joint Automated Repository for Various Integrated Simulations (JARVIS)[21] are prominent examples. These platforms use high-throughput first-principles calculations to evaluate various properties for a wide range of materials (stoichiometric and defect-free) and make the data publicly available. However, in the construction of these databases, the candidate materials for calculation are selected among known structures or based on known structural prototypes. Such knowledge is not evenly distributed: most literature only reports compounds with "good" properties in the aspect of interest[22], compounds that are structurally simple to enable property simulation, or compounds that are readily synthesizable and stable to facilitate property measurement; lower symmetry structures are less explored than higher symmetry ones. These, among other factors, lead to bias in the materials data platforms.

Data bias, a ubiquitous issue in data science, has been more recognized in the social science domain[23,24] but is often overlooked in physical sciences, including materials science. Just as it

causes social unfairness in social data, bias in materials data is harmful to data-driven materials modeling and design. An example is a bias in stability data among crystal structures, which we refer to as "structure–stability bias". Such bias hinders the modeling of phase stabilities, thus affecting the accurate prediction of microstructure. As Molkeri et al.[25] found, microstructure information is important for the modeling of various materials properties, therefore, the impact of structure–stability bias is not limited to stability itself but also on other properties.

Although some attempts have been pursued to characterize bias on trained models *post facto*[26] or reduce the impact of data bias on model training[27], few have addressed bias intrinsic to the data for which the models are trained and mitigated bias from the onset. The presence of bias in materials data may be inevitable since the distributions of properties are unknown and can be uneven in nature. Nonetheless, detecting the bias of datasets could alert users of their potential impact. As bias originates from uneven coverage of different materials families, it can be captured by examining the diversities of families in the data, which reflects the completeness of coverage. Moreover, by adding well-selected new data points, bias in a dataset can be reduced. Towards this end, the active learning method provides a way to select optimal data points guided by sampling strategies considering uncertainty, diversity, or performance, and has been applied to materials discovery[28-31]. With a specially designed sampling strategy, active learning can serve as a method for bias reduction.

In this work, we propose entropy-targeted active learning (ET-AL) as a systematic approach to detecting and reducing materials data bias. We focus on the structure–stability bias in DFT-generated databases as a use case for demonstrating the approach. With information entropy as a diversity metric, we quantify the bias of stability by its diversity among structures. We then develop an active learning method with a sampling strategy towards increasing the diversity of

stability of underrepresented structures, thus reducing the bias. We demonstrate the capability of ET-AL through experiments performed on existing datasets. We show that ET-AL provides a general method for mitigating bias in materials datasets and is also applicable in guiding the construction of materials databases, thus granting materials researchers access to low-bias data for machine learning.

## RESULTS AND DISCUSSION

### Data Bias Characterization

For demonstration purposes, we retrieve two materials datasets: (1) structure and formation energy per atom of all binary intermetallic compounds among the elements Al, Ti, Cr, Fe, Co, Ni, Cu, and W from OQMD (denoted OQMD-8, size 2,953); and (2) all entries with elastic moduli available from the JARVIS classical force-field inspired descriptors (CFID) dataset[32], cleaned as described in the Methods (denoted J-CFID, size 10,898). We show the distribution of formation energy per atom $\Delta E$ of materials in the two datasets with respect to crystal system in Figure 1a–b. We consider compounds in the cubic, hexagonal, trigonal, tetragonal, and orthorhombic systems to be higher in symmetry than those of monoclinic or triclinic systems, because they possess one or more rotation axes, and their unit cells have three or fewer free interaxial angles and lattice parameters. Among the seven crystal systems, the lower symmetry monoclinic and triclinic systems display higher distribution density in the more stable (lower $\Delta E$) region. This observation contradicts the empirical rules that materials with higher symmetry generally have higher stability: materials with higher symmetry are usually more close-packed and have higher coordination numbers, both of which lead to higher stability[33,34]. Such phenomenon is due to the imbalanced coverage of different crystal systems in the materials datasets, and we refer to this as "structure–stability bias".

Without assuming any prior knowledge such as the correlation between symmetry and stability, are we still able to capture the bias? To that end, we first define the diversity of a dataset by recognizing that for values of a continuous variable $Y$, the diversity can be quantified by information entropy[35]

$$h(Y) = \mathrm{E}[-\log f(y)] = -\int f(y) \log f(y) \, dy \tag{1}$$

where $f(y)$ is the underlying probability density function of $Y$. Note the difference between diversity and other characteristics such as uncertainty and variability: (1) Whereas uncertainty describes the state of random variables with incomplete or unknown information, diversity is an attribute of a deterministic dataset; (2) whereas variability emphasizes the difference between $Y$ values, diversity favors a more even coverage of $Y$ values. In general, we can group the data into clusters by any appropriate criterion and estimate $h(Y)$ for every cluster from the $Y$ values in the dataset, thus quantifying the diversity of $Y$ in every cluster. Based on the observation from Figure 1a–b, the OQMD-8 dataset has coverage of materials with diverse $\Delta E$ values in the high symmetry crystal systems, while the $\Delta E$ values in the triclinic and monoclinic systems are not diverse. The J-CFID dataset, on the other hand, lacks diversity in the high-symmetry crystal systems.
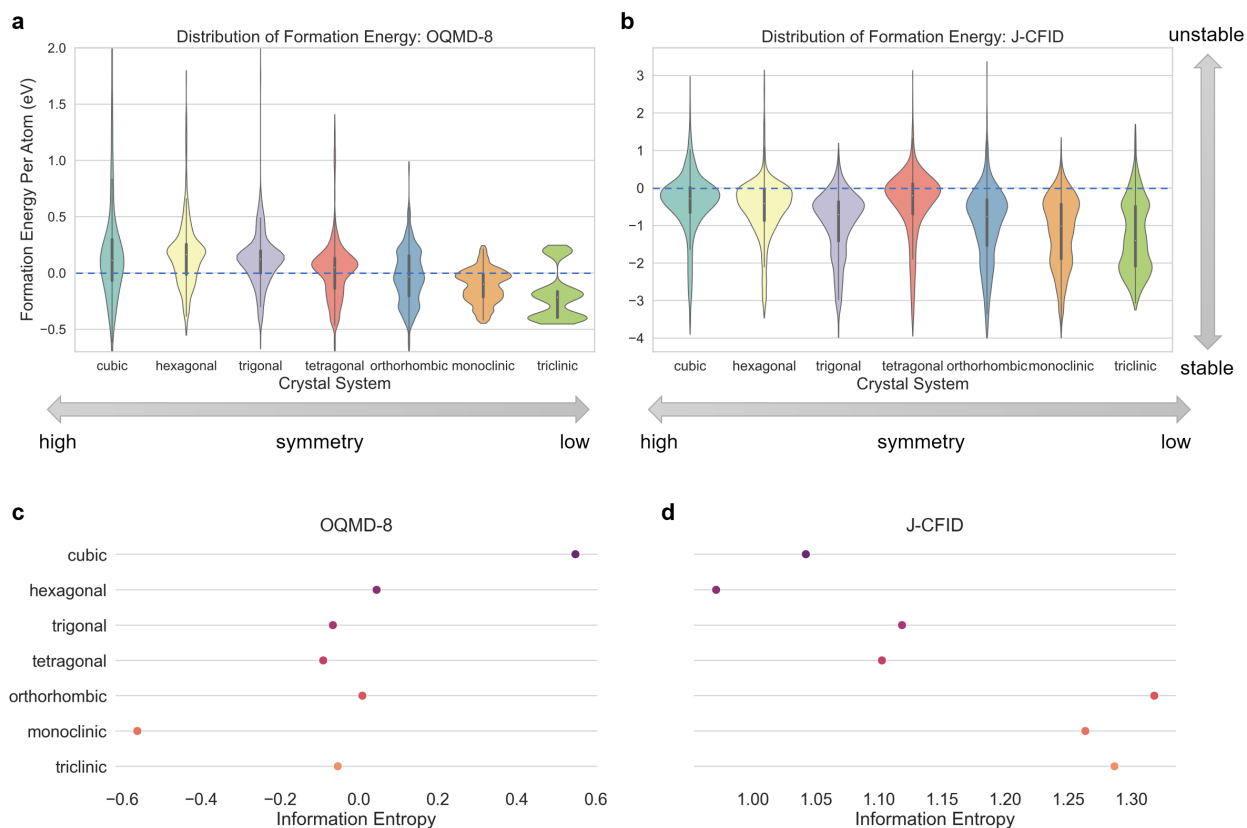
**Figure 1 Structure–stability bias in two datasets. a–b**, Kernel density estimation of the distribution of formation energy among different crystal systems in the OQMD-8 and J-CFID datasets. **c–d**, Information entropy of formation energy among different crystal systems in OQMD-8 and J-CFID datasets. Colors indicate the degrees of symmetry of crystal systems.

Next, we measure bias using a fairness criterion[36], i.e., the difference in $h(Y)$ between different clusters indicate the existence and level of bias. For our application, we will use crystal systems as natural clusters, and quantify the structure–stability bias via fairness of $h(\Delta E)$. Figure 1c–d shows that $h(\Delta E)$ captures the observed difference in diversities, thus reflecting the structure–stability bias.

## Active Learning for Bias Mitigation

With fairness in diversity as a measure, the data bias can be reduced systematically by adding data to the least diverse crystal system in a manner that increases its diversity in $\Delta E$. We develop the entropy-targeted active learning (ET-AL) algorithm (Figure 2) to attain this automatically. In the

active learning context, we refer to the materials with properties known and unknown as "labeled" and "unlabeled", respectively. The ET-AL algorithm iteratively picks a target crystal system (usually the least diverse one), selects an optimal unlabeled material that may improve $h(\Delta E)$ of the system and adds it to the labeled data. The iteration terminates when a pre-specified criterion is satisfied, or all materials are labeled. Details of the algorithm and its implementation are provided in Methods and Algorithm S 1.
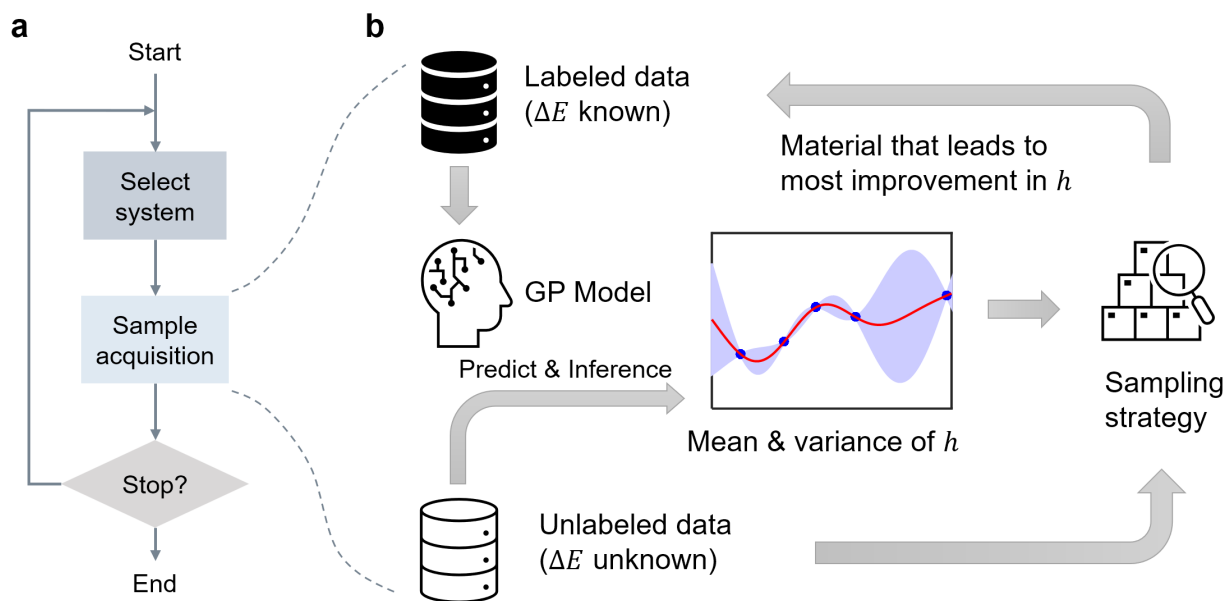


**Figure 2 Schematic of the ET-AL algorithm for data bias mitigation. a**, Overall procedure of ET-AL: a target crystal system is selected, then an unlabeled material is selected and labeled. The steps repeat until the stopping criteria are satisfied. **b**, The procedure of sample acquisition: a Gaussian Process (GP) model is trained with the labeled data and makes predictions for the unlabeled data. The predictive mean and variance of $h$ resulting from adding each material are inferred therefrom. Based on these, the optimal material is selected according to the sampling strategy and added to the labeled data.

## Experimentation and Demonstration

As a demonstration of the ET-AL method, we conduct experiments on the J-CFID dataset. The overall procedure is illustrated in Figure 3a: we split the dataset into a test set, a labeled set with artificial bias, and an unlabeled set. We use ET-AL to augment the labeled set into a low-bias training set (marked ETAL) and create another training set (marked RAND) of the same size by

randomly sampling from the unlabeled set. In addition to demonstrating that ET-AL effectively reduces the structure–stability bias, we show the impact such bias has by comparing supervised ML models for bulk modulus $B$ and shear modulus $G$ derived from the two training sets.
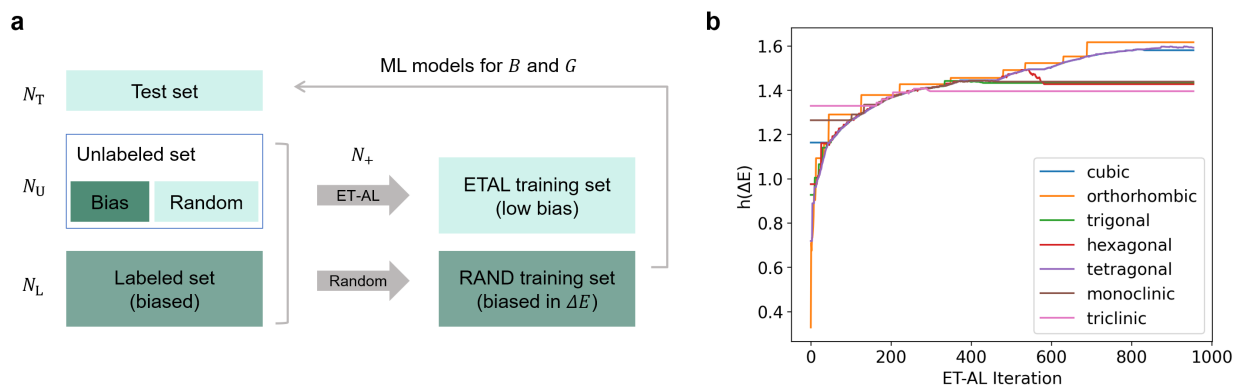


**Figure 3 Experiments on the J-CFID dataset. a**, Split of the dataset: $N_T$ entries are left out as the test set. From the remaining data, some entries are taken away to create an artificial bias and put into the unlabeled set together with randomly selected entries, in total $N_U$. The $N_T$ entries remaining form a labeled set with significant bias. Two training sets are constructed by adding the same number of samples from the unlabeled set to the labeled set, guided by ET-AL and randomly, respectively. **b**, Change of information entropy in every crystal system during ET-AL iteration.

In the experiment, we set $N_L = 1,000$, $N_U = 4,898$, and $N_T = 5,000$. The artificial bias is introduced by removing all tetragonal and trigonal materials with $\Delta E > 0$ and all orthorhombic materials with $\Delta E < 0$. With materials represented by graph embeddings (presented in Methods), ET-AL is applied to the dataset and runs for 985 iterations before termination. As Figure 3b shows, the introduced bias is captured by the diversity metric (the three manipulated crystal systems have relatively low initial $h$), and mitigated by ET-AL. Moreover, through ET-AL, the dataset reaches a state where diversities of crystal systems are closer to each other as compared to the initial state, which is favored by the fairness criterion (also shown in Figure S 1). A similar demonstration performed on the OQMD-8 dataset is presented in Figure S 2 and Figure S 3.

Next, we investigate the effects of ET-AL on dataset distribution. We employ t-distributed stochastic neighbor embedding (t-SNE)[37] for dimension reduction of the graph embedding

representations of J-CFID data into a 2-dimensional space. The low-dimensional embeddings acquired by t-SNE reflect the distribution of data in the structure space. In Figure 4, we use these embeddings to show the coverage of the labeled dataset (a; see also Figure S 4) and the ET-AL-selected and randomly selected data (b–c). ET-AL guides sampling in the underrepresented regions (where color is lighter in a), as opposed to a nearly uniform coverage by random sampling in c.
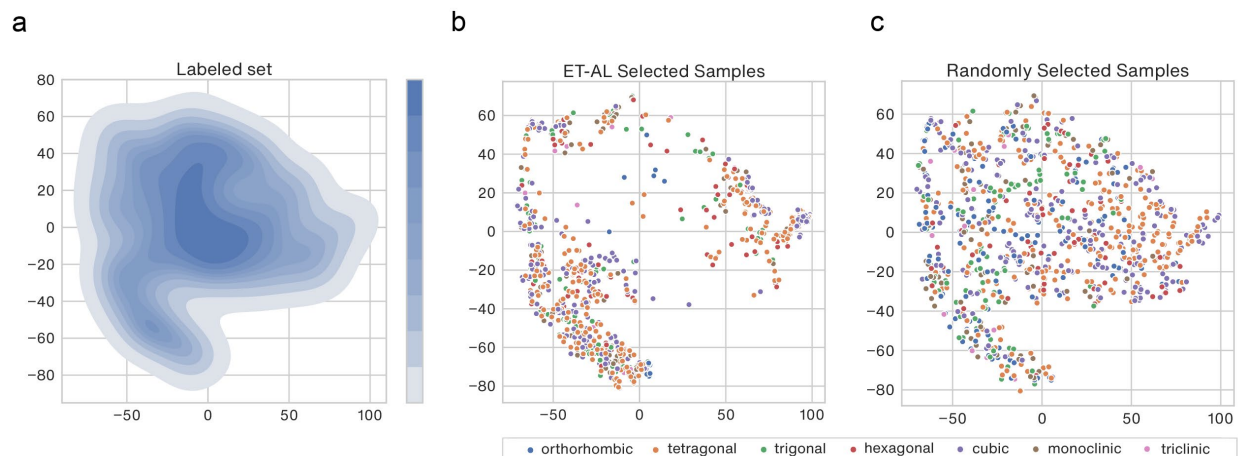


**Figure 4 Visualization of dataset distributions. a**, Kernel density estimation (KDE) plot of tSNE embeddings of the labeled dataset. The shade shows the density of points, regions with lighter colors are less covered. **b–c**, t-SNE plots of graph embeddings of the materials selected by ET-AL and random sampling, respectively, with colors indicating crystal systems. Compared to random sampling, materials selected by ET-AL better cover the region where labeled samples are sparse, as well as the crystal systems where artificial bias is introduced.

To assess the impact of bias on property prediction, we train multiple supervised learning models on the two training sets, both of size 1,954, for predicting $B$ and $G$ from a set of descriptors (detailed in Methods). Each model is trained 30 times with different random states (controlling the initialization, feature permutation, etc., but not affecting training data), and the coefficient of determination ($R^2$) on the test set is recorded. Models include random forest (RF), gradient boosting (GB), neural network (NN), and support vector regression (SVR), among which RF and GB attain relatively better performances on the task. A potential reason for such performance difference is that the descriptors form heterogeneous tabular data, for which tree ensemble models

have an advantage[38]. We summarize the performances of these ML models in Figure 5, from which we find that models derived from the ETAL dataset with reduced bias display systematically superior accuracies over those from the RAND dataset.

**a**                                                    **b**
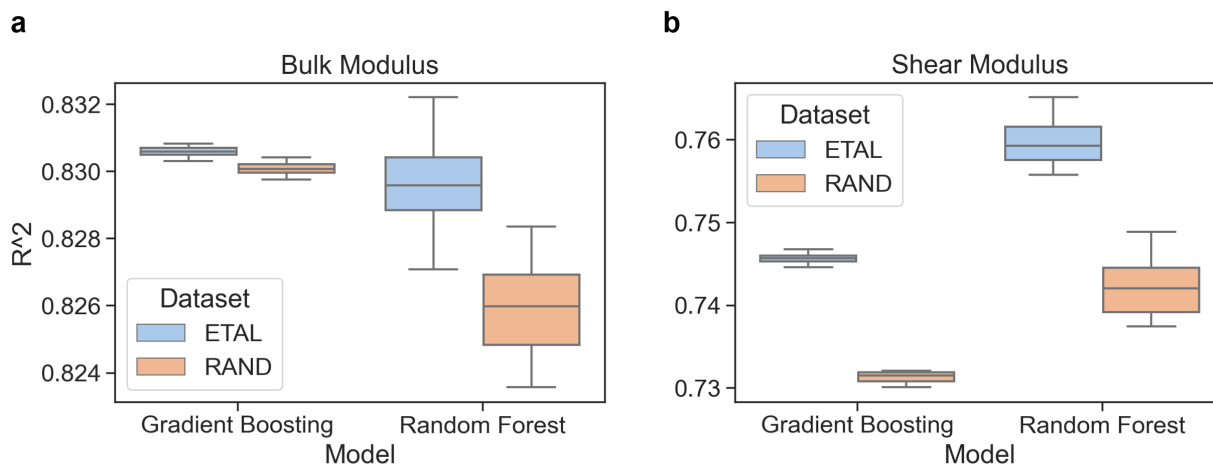


**Figure 5 Comparison of supervised ML models trained on two J-CFID-derived datasets.** The boxplots show testing $R^2$ of each model type and target property across 30 replicates, **a**, for bulk modulus and **b**, for shear modulus. Models trained on the ETAL dataset display higher testing $R^2$ than those trained on the RAND dataset.

In Figure 6, we mark the "most improved samples", i.e., testing samples for which the ML models' prediction accuracies using the ETAL training set show greater advantages compared to using the RAND training set. As observed, most of these samples are in the underrepresented regions of the labeled set (low-density regions in Figure 4a). ET-AL's focus in these regions during sample selection (triangles in Figure 6 overlap with sampling points in Figure 4b) leads to the better accuracy of ML models trained on the ETAL dataset. These observations agree with the findings of Li et al.[39]: ML models trained on a biased dataset lack generalizability to underrepresented test samples. ET-AL provides a solution to the problem by reducing structure–stability bias, which improves the coverage of the dataset in the structure space, and thus facilitates downstream tasks such as ML modeling of mechanical properties $B$ and $G$.
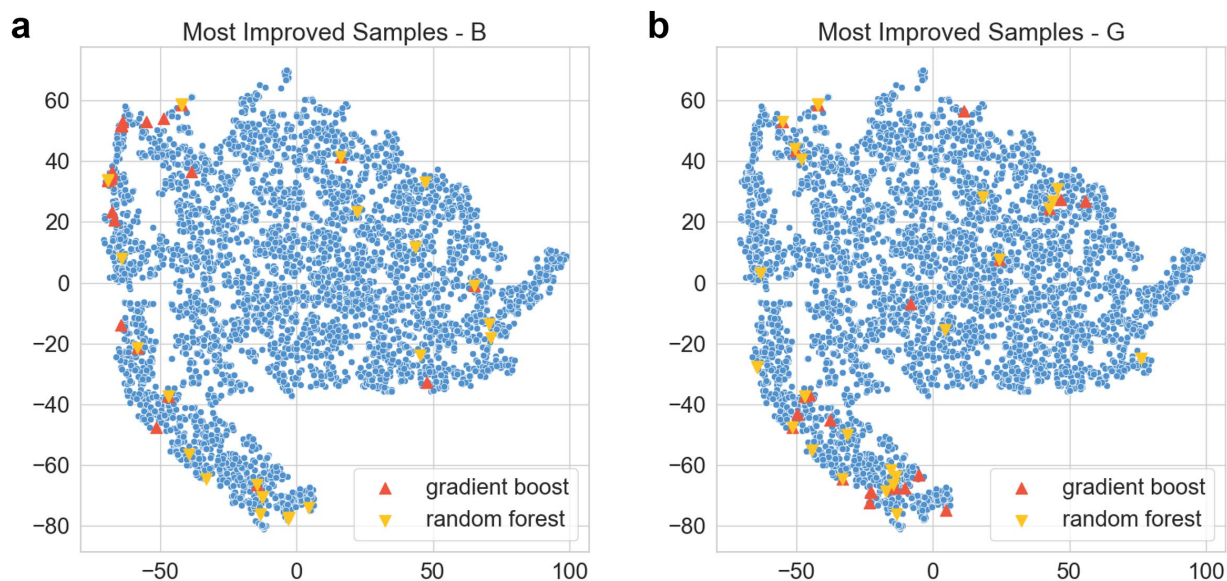
**Figure 6 Locations of the most improved samples.** Blue dots show the tSNE mapping of the test samples' graph embeddings. The triangles mark the samples where ML models trained on the ETAL set show the most advantages in accuracy over those trained on the RAND set, for **a**, bulk modulus and **b**, shear modulus. 30 samples are shown for each model–property combination.

## Potential Applications

The data bias metric and ET-AL method proposed in this work have a wide range of applications in materials discovery and beyond. First, researchers may examine and potentially reduce the bias in their datasets before developing data-driven models thereon. Second, ET-AL allows steering autonomous data acquisition in an unbiased way. This includes high-throughput computation, as well as experiments such as self-driving laboratories[40]. Though we presented the structure–stability bias as an example, the method applies to other forms of bias as well.

An application of particular significance is dealing with bias in materials data resources. Since new materials are continually added to the databases, ET-AL can fit in the pipeline to select the materials to add. In practice, however, some databases are so large that an observable effect of bias mitigation requires adding many new data points, and there are other considerations besides bias in database construction. In remediation, ET-AL can also guide trimming rather than

expanding a database, i.e., selecting a less biased subset. The level of bias can be tuned according to the need of usage.

Though originally proposed for materials data, ET-AL is generally applicable to other scientific and engineering domains, for example, data-intensive physics such as high-energy experiments[41] and observational astronomy[42], as well as cheminformatics[43] and bioinformatics[44], where data bias may lead to inaccuracies in parameter calibration, ML modeling, or diagnostics.

## CONCLUSION

We highlighted the previously overlooked bias in materials data resources, which has an impact on a broad range of data-driven materials modeling and design studies. We proposed a generic metric for data bias based on diversity measured by information entropy, which successfully captures the structure–stability bias in datasets retrieved from widely used materials data platforms OQMD and JARVIS. We then formulated and implemented an entropy-target active learning (ET-AL) framework to automatically reduce bias in datasets by acquiring new samples. Through ablation studies, we demonstrated that ET-AL can effectively reduce the structure–stability bias, thus improving data coverage in the structure space and increasing the accuracy of data-driven modeling of materials properties.

We also note that as a generic framework, ET-AL's capability is not limited to materials databases. As the data-driven research paradigm has been adopted by various domains, and data bias is ubiquitous in almost every data system, we anticipate that the ET-AL method is applicable to a variety of domains, such as physics, chemistry, and life science, to facilitate the curation of high-quality data and data-driven studies.

## METHODS

### Dataset Preparation

**Data collection and cleaning.** The OQMD-8 dataset is retrieved from OQMD using its API implemented in the "qmpy-rester" package. The J-CFID dataset is downloaded from figshare.com[32]. Out of the >50,000 entries, the ones reporting positive $B$ and $G$ values are kept. Entries containing elements H, Tc, halogens (VIIA), noble gases (VIIIA), lanthanum family, and those with atomic numbers ≥84 are excluded. Figure S 5 and Figure S 6 show some statistics of the J-CFID dataset.

**Graph embedding representation.** We use the crystal graph convolutional neural network (CGCNN) model[45], which maps a graph encoding of crystal structure to a numerical vector before predicting its properties. The numerical vector provides a representation of materials' structures regulated by the target property. We feed the crystallographic information framework (CIF) files of all entries in the dataset to a pretrained CGCNN model and extract a 32-dimensional vector representation for each material's structure. Unless specified otherwise, we use the model pretrained on formation energy per atom.

### Information Entropy

Information entropy of continuous-valued $\Delta E$ is estimated from a discrete set of $\Delta E$ values using the "differential_entropy" function from the scipy.stats package[46]. The software automatically selects a numerical method for entropy estimation[47] based on data size; details of the numerical estimation methods are described in Supplemental Information (SI).

### Active Learning

**Target system selection.** In every iteration, all crystal systems with unlabeled sample(s) available are candidates. Systems that are sampled but not improved five consecutive times are excluded. Of the remaining candidates, the crystal system with the lowest $h(\Delta E)$ is selected as the target.

**Gaussian process modeling.** Gaussian Process (GP) modeling[48] builds upon the assumption that the responses $\boldsymbol{y} = \{y_1, \dots, y_m\}$ are jointly Gaussian distributed given the predictors $\boldsymbol{X} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_m\}$:

$$\boldsymbol{y}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{K} + \sigma^2 \boldsymbol{I}) \tag{2}$$

where the covariance matrix $\boldsymbol{K}$ is inferred from the similarity between predictors using a kernel function, and the $\sigma^2 \boldsymbol{I}$ term accounts for noise. Once trained, inputting an unseen predictor $\hat{\boldsymbol{x}}$, the model outputs not a single value but a predicted Gaussian distribution of the response $\hat{y}$. Hence, GP is an uncertainty-aware machine learning model. We train GP models using the "ExactGPModel" module of GPyTorch package[49], with the graph embedding representation as predictors and $\Delta E$ as response.

**Monte Carlo inference.** For each unlabeled material, the GP model provides a predicted distribution of $\Delta E$, from which we use the Monte Carlo method[50] to infer the resulting change in $h(\Delta E)$ by adding the material, as illustrated in Figure 7. We thereby obtain the predictive mean and variance of $h$ for every unlabeled material, which are later used in the evaluation of the sampling criterion.
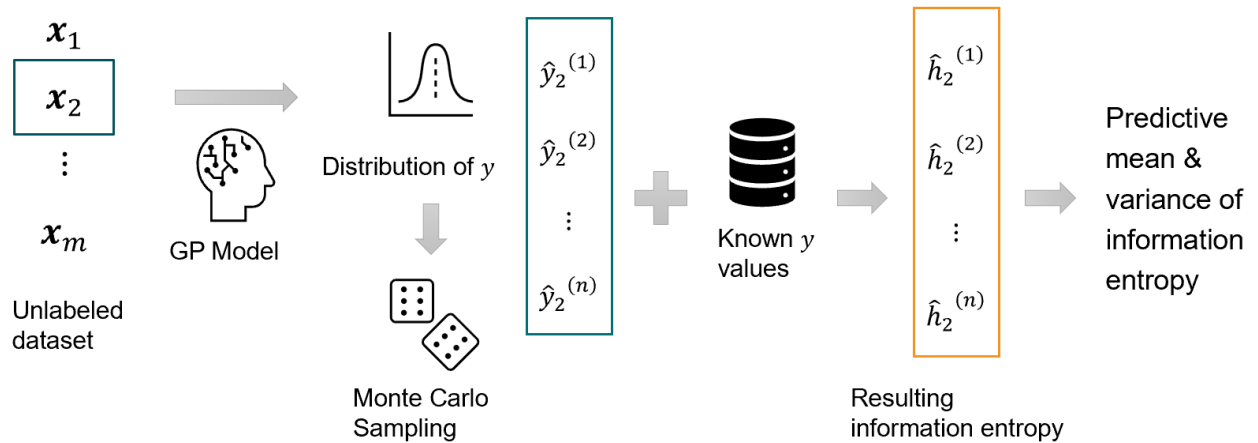


**Figure 7 Schematic of the Monte Carlo Inference for mean and variance of $h$.** For simplicity, formation energy or any property of interest is denoted as $y$ in the figure. For each unlabeled material, we obtain a predicted distribution of $y$ via the GP model, and randomly draw $n$ samples therefrom. By trying to add

every sample into the $y$ values of labeled dataset, we get $n$ resulting $h$ values ($n = 1,000$ is used in experiments), from which we can obtain the predictive mean and variance of the resulting $h$ if that material is selected.

**Sampling strategy.** The selection of unlabeled materials is guided by the expected improvement (EI) sampling criterion[51]. For every unlabeled material $x$, the expected improvement in $h$

$$\text{EI}(x) = \text{E}[\max\{0, \Delta(x)\}] = \hat{s}(x)\phi\left(\frac{\Delta(x)}{\hat{s}(x)}\right) + \Delta(x)\Phi\left(\frac{\Delta(x)}{\hat{s}(x)}\right) \tag{3}$$

is evaluated, where $\Delta(x) = h(x) - h_{\text{cur}}$ is the difference between the predicted mean $h$ and the current $h$; $\hat{s}(x)$ is the predicted standard deviation of $h$; $\phi(\cdot)$ and $\Phi(\cdot)$ are the probability density function (*pdf*) and cumulative distribution function (*cdf*) of standard Gaussian distribution, respectively. The unlabeled material with the largest EI is selected, i.e., $x^* = \arg\max_x \text{EI}(x)$.

**Stopping criteria.** In our experiments on the J-CFID dataset, the active learning process is terminated when the target crystal system has the highest $h(\Delta E)$ of all systems. In that case, improving its $h$ will worsen the fairness of diversity. However, when evaluation (e.g., first-principles calculation) of new materials is feasible, the existence of materials improving $h$ of the least diverse system is almost guaranteed. In such application scenarios, stopping criteria may be specified according to resources and budget, or may not be needed.

## Supervised Machine Learning

**Materials featurization.** Compositional and structural descriptors are retrieved for every crystal structure using an automatic workflow developed by Georgescu et al.[52], built upon the pymatgen[53] and Matminer[54] toolkits. Descriptors that contain NaN values or have zero variance among the dataset are removed, resulting in 117 descriptors for every material.

**Machine learning.** The supervised ML models are trained and tested using implementations in the scikit-learn package[55] with default settings unless specified otherwise. For random forests, the

maximum depth of trees is not limited. For gradient boosting, the maximum depth of individual trees is 3. For SVM, the radial basis function (RBF) kernel is used. For neural networks, model architecture and hyperparameters are tuned using a 5-fold cross-validated grid search. The configuration used is 1 hidden layer with 128 neurons, $L_2$ regularization strength $\alpha = 0.1$, rectified linear unit (ReLU) activation for hidden layers, and Adam optimizer.

## Resource Availability

### Lead contact

Further information and requests for resources and materials should be directed to and will be fulfilled by the lead contact, Professor Wei Chen (weichen@northwestern.edu).

### Materials availability

This study did not generate new, unique reagents.

### Data and code availability

The datasets used in bias characterization and experimentation are publicly available: the OQMD-8 dataset was obtained from oqmd.org, and the J-CFID dataset was obtained from figshare.com[32]. The code and data generated in the current study are available on the ET-AL GitHub repository: https://github.com/Henrium/ET-AL, and on Zenodo through the DOI: 10.5281/zenodo.7406332.

## AUTHOR CONTRIBUTIONS

## DECLARATION OF INTERESTS

The authors declare no competing interests.

## REFERENCES

1. Chen, W., Iyer, A., and Bostanabad, R. (2022). Data Centric Design: A New Approach to Design of Microstructural Material Systems. Engineering *10*, 89-98. 10.1016/j.eng.2021.05.022.
2. Himanen, L., Geurts, A., Foster, A.S., and Rinke, P. (2019). Data‑Driven Materials Science: Status, Challenges, and Perspectives. Advanced Science *6*, 1900808. 10.1002/advs.201900808.
3. Stein, H.S., and Gregoire, J.M. (2019). Progress and prospects for accelerating materials science with automated and autonomous workflows. Chemical Science *10*, 9640-9649. 10.1039/c9sc03766g.
4. Agrawal, A., and Choudhary, A. (2016). Perspective: Materials informatics and big data: Realization of the "fourth paradigm" of science in materials science. APL Materials *4*, 053208. 10.1063/1.4946894.
5. Umehara, M., Stein, H.S., Guevarra, D., Newhouse, P.F., Boyd, D.A., and Gregoire, J.M. (2019). Analyzing machine learning models to accelerate generation of fundamental materials insights. npj Computational Materials *5*, 34. 10.1038/s41524-019-0172-5.
6. Gong, S., Wang, S., Zhu, T., Chen, X., Yang, Z., Buehler, M.J., Shao-Horn, Y., and Grossman, J.C. (2021). Screening and Understanding Li Adsorption on Two-Dimensional Metallic Materials by Learning Physics and Physics-Simplified Learning. JACS Au *1*, 1904-1914. 10.1021/jacsau.1c00260.

7.  Oganov, A.R., Pickard, C.J., Zhu, Q., and Needs, R.J. (2019). Structure prediction drives materials discovery. Nature Reviews Materials. 10.1038/s41578-019-0101-8.

8.  Ye, W., Chen, C., Wang, Z., Chu, I.-H., and Ong, S.P. (2018). Deep neural networks for accurate predictions of crystal stability. Nature Communications *9*, 3800. 10.1038/s41467-018-06322-x.

9.  Choudhary, K., and Decost, B. (2021). Atomistic Line Graph Neural Network for improved materials property predictions. npj Computational Materials *7*. 10.1038/s41524-021-00650-1.

10. Tian, Y., Xue, D., Yuan, R., Zhou, Y., Ding, X., Sun, J., and Lookman, T. (2021). Efficient estimation of material property curves and surfaces via active learning. Physical Review Materials *5*. 10.1103/physrevmaterials.5.013802.

11. Arróyave, R., and McDowell, D.L. (2019). Systems Approaches to Materials Design: Past, Present, and Future. Annual Review of Materials Research *49*, 103-126. 10.1146/annurev-matsci-070218-125955.

12. Wang, Y., Iyer, A., Chen, W., and Rondinelli, J.M. (2020). Featureless adaptive optimization accelerates functional electronic materials design. Applied Physics Reviews *7*, 041403. 10.1063/5.0018811.

13. Yao, Z., Sánchez-Lengeling, B., Bobbitt, N.S., Bucior, B.J., Kumar, S.G.H., Collins, S.P., Burns, T., Woo, T.K., Farha, O.K., Snurr, R.Q., and Aspuru-Guzik, A. (2021). Inverse design of nanoporous crystalline reticular materials with deep generative models. Nature Machine Intelligence *3*, 76-86. 10.1038/s42256-020-00271-1.

14. Mansouri Tehrani, A., Oliynyk, A.O., Parry, M., Rizvi, Z., Couper, S., Lin, F., Miyagi, L., Sparks, T.D., and Brgoch, J. (2018). Machine Learning Directed Search for Ultraincompressible, Superhard Materials. Journal of the American Chemical Society *140*, 9844-9853. 10.1021/jacs.8b02717.

15. Sambasivan, N., Kapania, S., Highfill, H., Akrong, D., Paritosh, P., and Aroyo, L.M. (2021). "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, 2021-05-06. (ACM).

16. Liang, W., Tadesse, G.A., Ho, D., Li, F.-F., Zaharia, M., Zhang, C., and Zou, J. (2022). Advances, challenges and opportunities in creating data for trustworthy AI. Nature Machine Intelligence. 10.1038/s42256-022-00516-1.

17. de Pablo, J.J., Jackson, N.E., Webb, M.A., Chen, L.-Q., Moore, J.E., Morgan, D., Jacobs, R., Pollock, T., Schlom, D.G., Toberer, E.S., et al. (2019). New frontiers for the materials genome initiative. npj Computational Materials *5*, 41. 10.1038/s41524-019-0173-4.

18. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K.A. (2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL Materials *1*, 011002. 10.1063/1.4812323.

19. Saal, J.E., Kirklin, S., Aykol, M., Meredig, B., and Wolverton, C. (2013). Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). JOM *65*, 1501-1509. 10.1007/s11837-013-0755-4.

20.     Kirklin, S., Saal, J.E., Meredig, B., Thompson, A., Doak, J.W., Aykol, M., Rühl, S., and Wolverton, C. (2015). The Open Quantum Materials Database (OQMD): assessing the accuracy of DFT formation energies. npj Computational Materials *1*, 15010. 10.1038/npjcompumats.2015.10.

21.     Choudhary, K., Garrity, K.F., Reid, A.C.E., Decost, B., Biacchi, A.J., Hight Walker, A.R., Trautt, Z., Hattrick-Simpers, J., Kusne, A.G., Centrone, A., et al. (2020). The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. npj Computational Materials *6*. 10.1038/s41524-020-00440-1.

22.     Jia, X., Lynch, A., Huang, Y., Danielson, M., Lang'At, I., Milder, A., Ruby, A.E., Wang, H., Friedler, S.A., Norquist, A.J., and Schrier, J. (2019). Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. Nature *573*, 251-255. 10.1038/s41586-019-1540-5.

23.     Celis, L.E., Keswani, V., and Vishnoi, N. (2020). Data preprocessing to mitigate bias: A maximum entropy based approach. Proceedings of the 37th International Conference on Machine Learning, (PMLR), pp. 1349–1359.

24.     Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. ACM Computing Surveys *54*, 1-35. 10.1145/3457607.

25.     Molkeri, A., Khatamsaz, D., Couperthwaite, R., James, J., Arróyave, R., Allaire, D., and Srivastava, A. (2022). On the importance of microstructure information in materials design: PSP vs PP. Acta Materialia *223*, 117471. 10.1016/j.actamat.2021.117471.

26.     McDuff, D., Ma, S., Song, Y., and Kapoor, A. (2019). Characterizing bias in classifiers using generative models. Advances in Neural Information Processing Systems, (Curran Associates, Inc.).

27.     Jiang, H., and Nachum, O. (2020). Identifying and correcting label bias in machine learning. (PMLR), pp. 702-712.

28.     Tran, K., and Ulissi, Z.W. (2018). Active learning across intermetallics to guide discovery of electrocatalysts for CO2 reduction and H2 evolution. Nature Catalysis *1*, 696-703. 10.1038/s41929-018-0142-1.

29.     Kusne, A.G., Yu, H., Wu, C., Zhang, H., Hattrick-Simpers, J., Decost, B., Sarker, S., Oses, C., Toher, C., Curtarolo, S., et al. (2020). On-the-fly closed-loop materials discovery via Bayesian active learning. Nature Communications *11*, 5966. 10.1038/s41467-020-19597-w.

30.     Ament, S., Amsler, M., Sutherland, D.R., Chang, M.-C., Guevarra, D., Connolly, A.B., Gregoire, J.M., Thompson, M.O., Gomes, C.P., and Dover, R.B.v. (2021). Autonomous materials synthesis via hierarchical active learning of nonequilibrium phase diagrams. Science Advances *7*, eabg4930. doi:10.1126/sciadv.abg4930.

31.     Yuan, R., Liu, Z., Balachandran, P.V., Xue, D., Zhou, Y., Ding, X., Sun, J., Xue, D., and Lookman, T. (2018). Accelerated Discovery of Large Electrostrains in BaTiO3‐Based Piezoelectrics Using Active Learning. Adv. Mater. *30*, 1702884. 10.1002/adma.201702884.

32.     Choudhary, K. (2021). JARVIS-DFT 3D dataset (cfid_33d-8-18-2021.json). https://figshare.com/articles/dataset/jdft_3d-7-7-2018_json/6815699.

33.    Pauling, L. (1929). The principles determining the structure of complex ionic crystals. Journal of the American Chemical Society *51*, 1010-1026. 10.1021/ja01379a006.

34.    Newnham, R.E. (2004). Symmetry. In Properties of Materials: Anisotropy, Symmetry, Structure, (Oxford University Press), pp. 14–22. 10.1093/oso/9780198520757.003.0005.

35.    Thomas, M., and Joy, A.T. (2006). Elements of information theory (Wiley-Interscience).

36.    Zhang, C., and Shah, J.A. (2014). Fairness in multi-agent sequential decision-making. Advances in Neural Information Processing Systems, (Curran Associates, Inc.).

37.    Maaten, L.v.d., and Hinton, G. (2008). Visualizing Data using t-SNE. Journal of Machine Learning Research *9*, 2579−2605.

38.    Grinsztajn, L., Oyallon, E., and Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on tabular data? Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track.

39.    Li, K., DeCost, B., Choudhary, K., Greenwood, M., and Hattrick-Simpers, J. (2022). A critical examination of robustness and generalizability of machine learning prediction of materials properties. arXiv preprint arXiv:2210.13597.

40.    Stach, E., Decost, B., Kusne, A.G., Hattrick-Simpers, J., Brown, K.A., Reyes, K.G., Schrier, J., Billinge, S., Buonassisi, T., Foster, I., et al. (2021). Autonomous experimentation systems for materials development: A community perspective. Matter *4*, 2702-2726. 10.1016/j.matt.2021.06.036.

41.    Gambhir, R., Nachman, B., and Thaler, J. (2022). Bias and priors in machine learning calibrations for high energy physics. Physical Review D *106*. 10.1103/physrevd.106.036011.

42.    Teerikorpi, P. (1997). Observational selection bias affecting the determination of the extragalactic distance scale. Annual Review of Astronomy and Astrophysics *35*, 101-136. 10.1146/annurev.astro.35.1.101.

43.    Sieg, J., Flachsenberg, F., and Rarey, M. (2019). In Need of Bias Control: Evaluating Chemical Data for Machine Learning in Structure-Based Virtual Screening. Journal of Chemical Information and Modeling *59*, 947-961. 10.1021/acs.jcim.8b00712.

44.    Zheng, W., Chung, L.M., and Zhao, H. (2011). Bias detection and correction in RNA-Sequencing data. BMC Bioinformatics *12*, 290. 10.1186/1471-2105-12-290.

45.    Xie, T., and Grossman, J.C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical Review Letters *120*, 145301. 10.1103/PhysRevLett.120.145301.

46.    Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods *17*, 261-272. 10.1038/s41592-019-0686-2.

47.    Alizadeh Noughabi, H. (2015). Entropy Estimation Using Numerical Methods. Annals of Data Science *2*, 231-241. 10.1007/s40745-015-0045-9.

48.    Rasmussen, C.E., and Williams, C.K.I. (2005). Gaussian Processes for Machine Learning (MIT Press).

49. Gardner, J., Pleiss, G., Weinberger, K.Q., Bindel, D., and Wilson, A.G. (2018). GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. Advances in Neural Information Processing Systems, (Curran Associates, Inc.).

50. Metropolis, N., and Ulam, S. (1949). The Monte Carlo Method. Journal of the American Statistical Association *44*, 335-341. 10.2307/2280232.

51. Jones, D.R., Schonlau, M., and Welch, W.J. (1998). Efficient Global Optimization of Expensive Black-Box Functions. Journal of Global Optimization *13*, 455-492. 10.1023/A:1008306431147.

52. Georgescu, A.B., Ren, P., Toland, A.R., Zhang, S., Miller, K.D., Apley, D.W., Olivetti, E.A., Wagner, N., and Rondinelli, J.M. (2021). Database, Features, and Machine Learning Model to Identify Thermally Driven Metal–Insulator Transition Compounds. Chemistry of Materials *33*, 5591-5605. 10.1021/acs.chemmater.1c00905.

53. Ong, S.P., Richards, W.D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V.L., Persson, K.A., and Ceder, G. (2013). Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. Computational Materials Science *68*, 314-319. 10.1016/j.commatsci.2012.10.028.

54. Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N.E.R., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M., et al. (2018). Matminer: An open source toolkit for materials data mining. Computational Materials Science *152*, 60-69. 10.1016/j.commatsci.2018.05.018.

55. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research *12*, 2825-2830.

# Supplemental Items

Algorithm S 1 Pseudocode of the ET-AL algorithm.

---

**Entropy-targeted active learning.**

**Input:** Labeled dataset $\mathcal{D} = \{x_i, y_i\}$, unlabeled dataset $\mathcal{U} = \{x_j'\}$; $\mathcal{D} = \cup_c \mathcal{D}_c$, $\mathcal{U} = \cup_c \mathcal{U}_c$ ($c \in$ crystal systems); Monte Carlo sample size $n$; stopping criteria

**while** stopping criteria are not satisfied **do**

    Calculate information entropies $H(\mathcal{D}_c) = h(y \in \mathcal{D}_c)$, select $c^* = \arg\min_c H(\mathcal{D}_c)$

    Fit a GP to $\mathcal{D}_{c*}$: $Y \sim GP(X)$

    **for** $x_j' \in \mathcal{U}_{c*}$ **do**

        Draw $n$ samples $\left\{y_j^{(k)}\right\}_{k=1}^n$ from $GP(x_j')$

        Calculate $h_j^{(k)} = H\left(\mathcal{D} \cup \left\{y_j^{(k)}\right\}\right)$ **for** $k = 1$ **to** $n$

        Calculate the mean and variance of $\left\{h_j^{(k)}\right\}_{k=1}^n$, $\mathrm{EI}(x_j')$ according to Equation (3)

    **end for**

    Select sample $x^* = \arg\max_{x' \in \mathcal{U}_{c*}} \mathrm{EI}(x')$

    Acquire $y^*$, remove $x^*$ from $\mathcal{U}_{c*}$, add $x^*, y^*$ to $\mathcal{D}_{c*}$

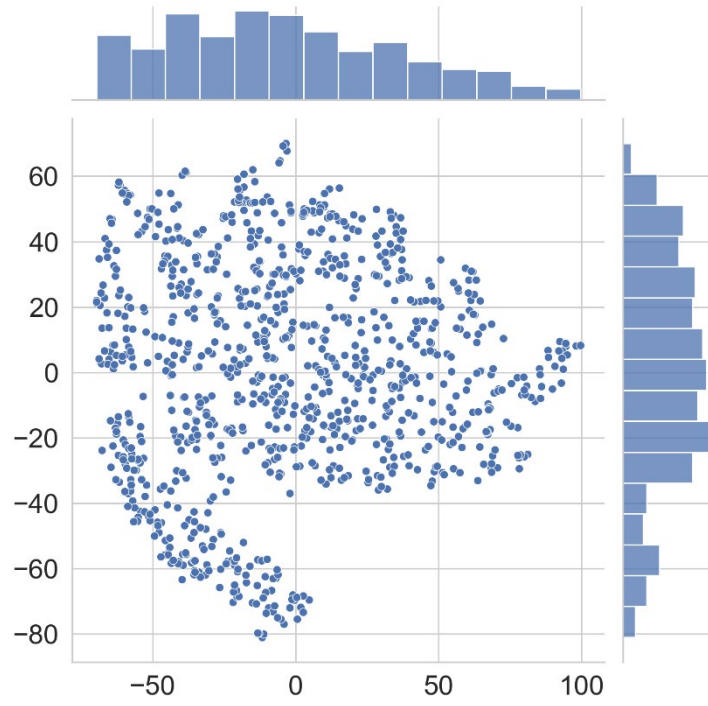**end while**

**Returns:** Updated dataset $\mathcal{D}$

---

Figure S 1 t-SNE plot of graph embeddings of the labeled dataset. The histograms show distribution densities in different regions. Related to Figure 4.
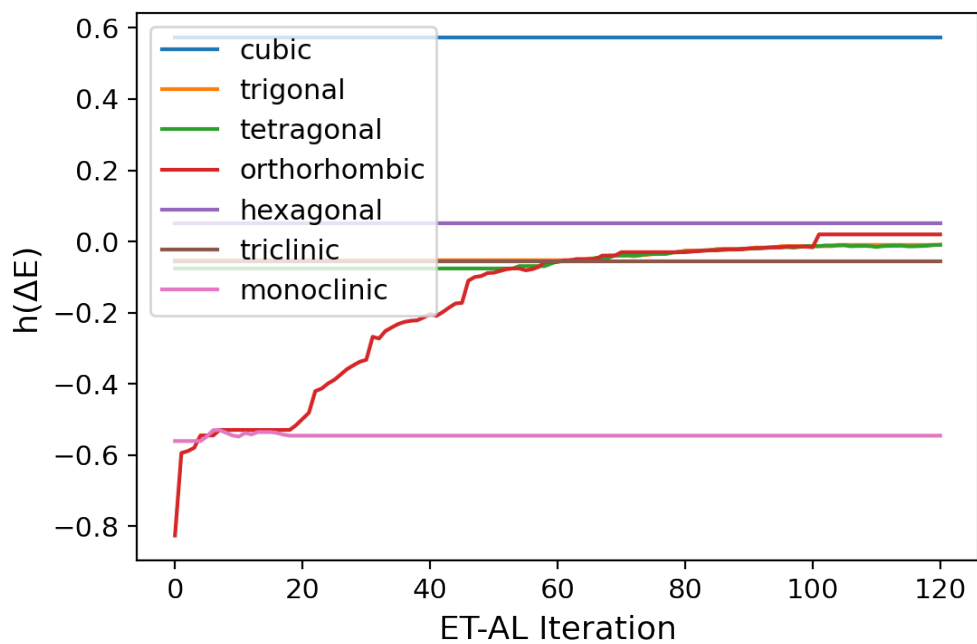
Figure S 2 Change of information entropy in every crystal system during ET-AL iteration on OQMD-8 dataset, related to Figure 3b. Artificial bias is introduced by leaving out all orthorhombic compounds with $\Delta E > 0$, together with randomly selected compounds, in total 1,000, as the unlabeled dataset. $h(\Delta E)$ for monoclinic is not improved in later iterations because new materials are selected only in the unlabeled dataset, which is limited by the diversity of OQMD-8 data.
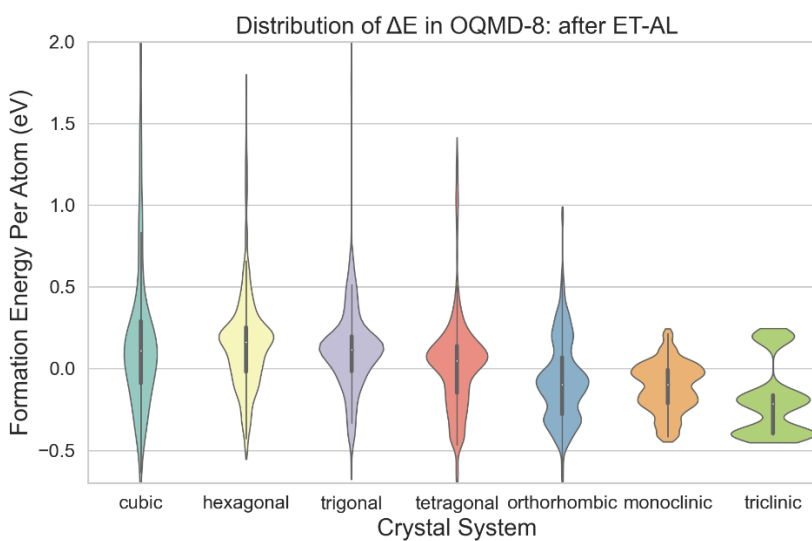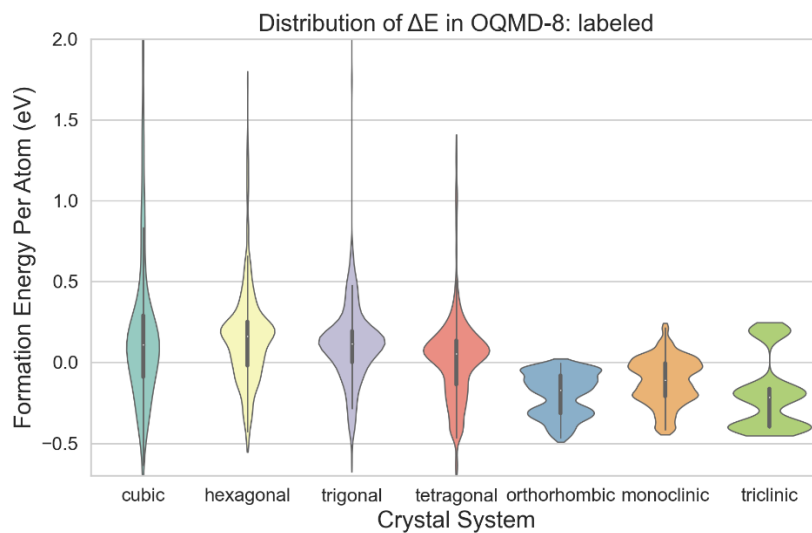
Figure S 3 Kernel density estimation of the distribution of $\Delta E$ among different crystal systems in the OQMD-8 derived datasets: (top) before and (bottom) after ET-AL.
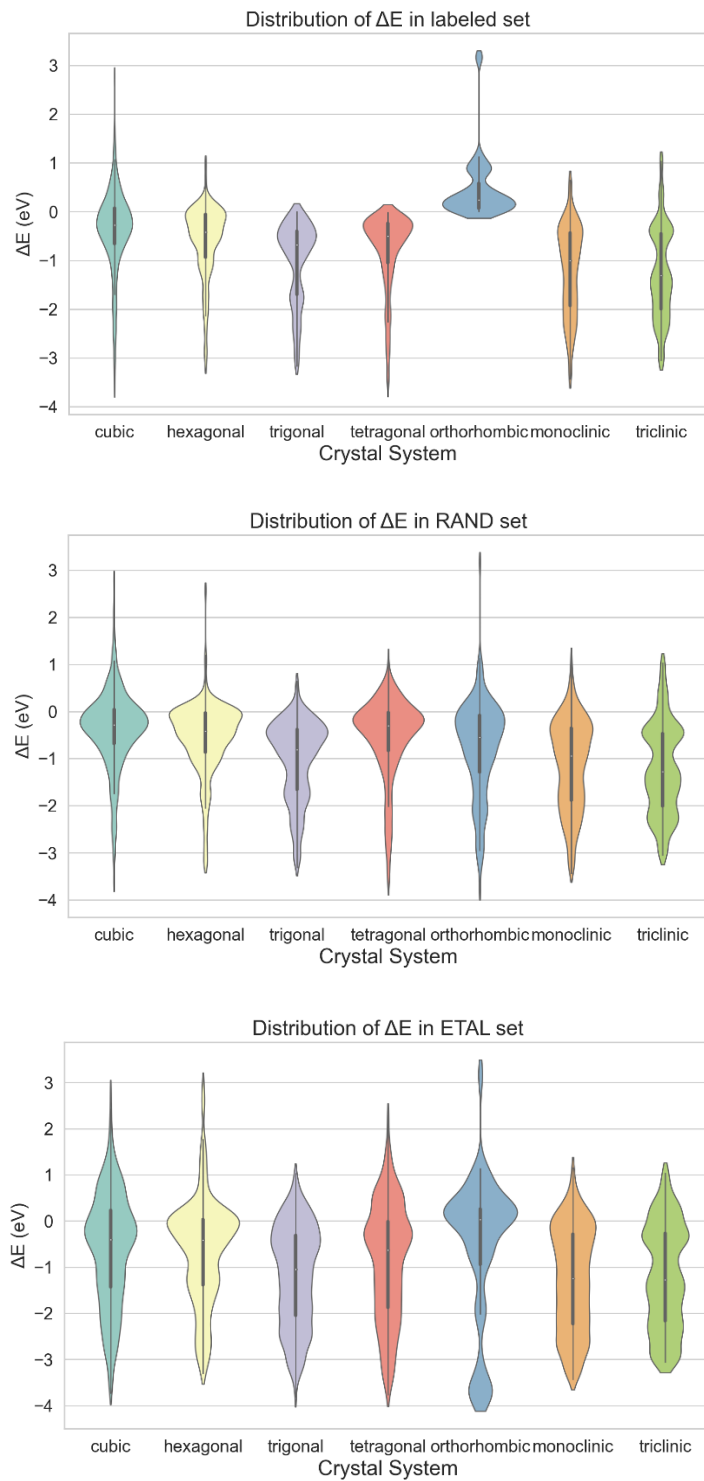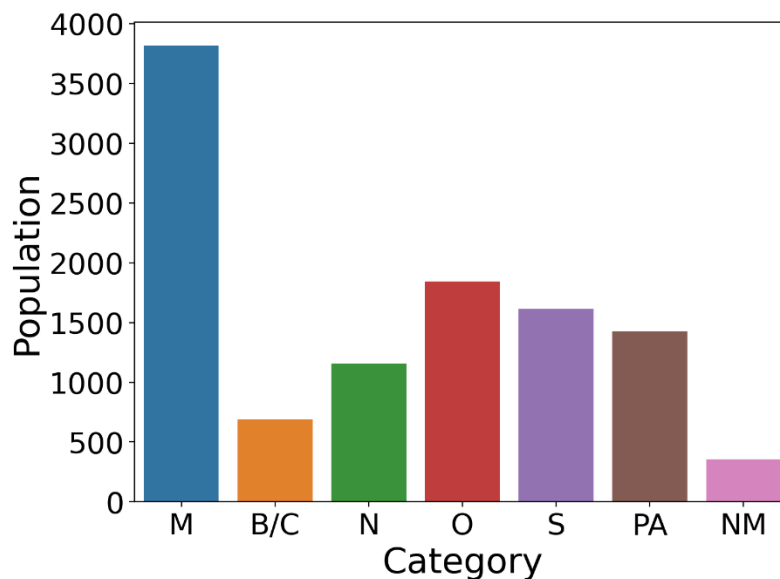
Figure S 4 Kernel density estimation of the distribution of $\Delta E$ among different crystal systems in the J-CFID-derived (from top to bottom) labeled, RAND, and ETAL datasets. Related to Figure 1b and Figure 3.

(a) The population of the J-CFID dataset by categories. M: metal or intermetallic. Columns 2–5 are compounds with metallic element(s) and a single type of anion: B/C: B/C/Si anion; N: N/P/As anion; O: O anion; S: S/Se/Te anion. PA: polyanionic, i.e., compounds with more than one type of anions above. NM: nonmetal, i.e., compounds without metallic elements.
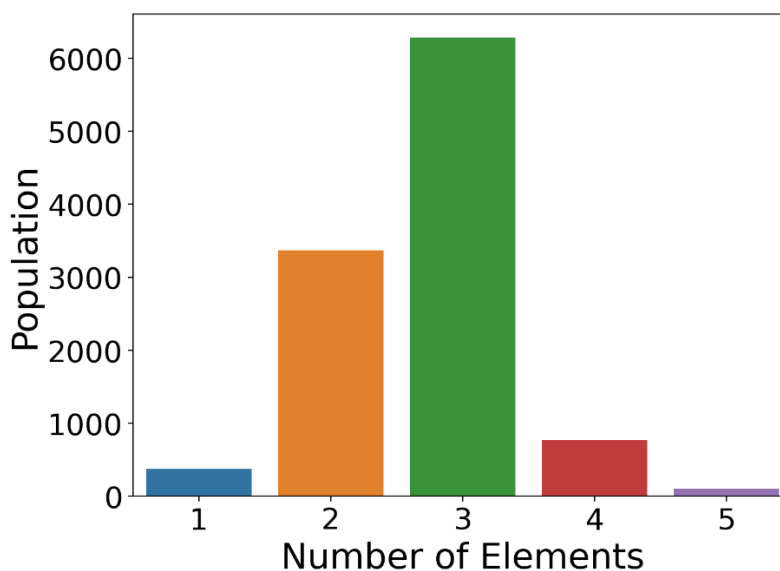


(b) The population of the J-CFID dataset by the number of component elements.

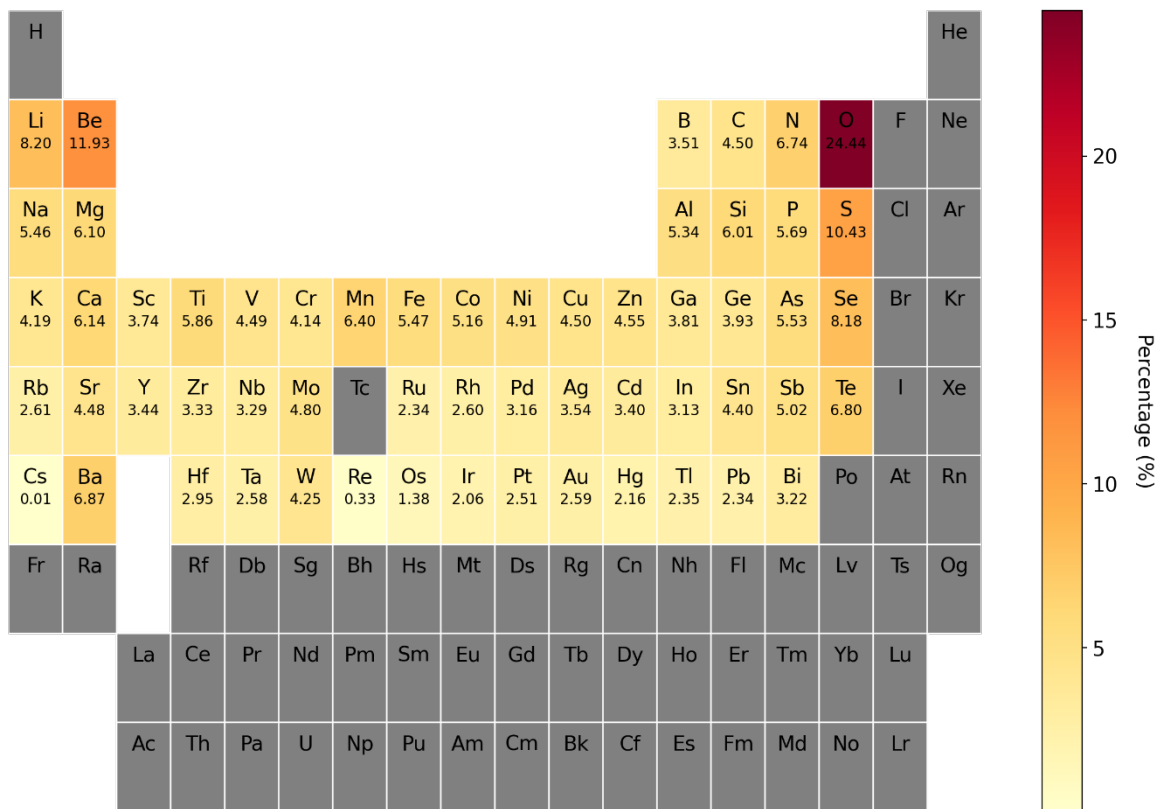Figure S 5 Statistics of the J-CFID dataset by number and types of component elements.

Figure S 6 Percentage of compounds containing each element in the J-CFID dataset.

# Supplemental Experimental Procedures

## 1. Bias metric

### Metric selection

One definition of data bias is "unjustifiable concentration on a particular part"[1]. In the context of this work, the "parts" can be regions in a space broadly defined by composition, (micro)structure, property, processing, or energy-based descriptions. As an example, the structure–stability bias arises from concentration (uneven coverage) in materials with certain structures (crystal systems) and stability ($\Delta E$). Such uneven coverage can be captured by the different diversities of stability among different crystal systems. The information entropy

$$h(Y) = -\int f(y) \log f(y) \mathrm{d}y$$

has been widely adopted as a metric for diversity[2]. Its numerical nature and simplicity in the calculation are desirable as a target in active learning. Also widely adopted as a numerical diversity measure is the determinantal point processes (DPP)[3]. However, DPP is evaluated pairwise, thus lacking scalability to large materials databases.

Another seemingly applicable metric is the conditional information entropy

$$h(Y|CS) = -\int f(cs, y) \log f(y|cs) \mathrm{d}cs \mathrm{d}y$$

where $CS$ denotes the crystal system. However, this metric has several problems. First, $f(cs, y) = f(y|cs) \cdot f(cs)$, where $f(cs)$, the probability distribution of $CS$, is related to the fractions of crystal systems. The "true" fractions in nature are unknown, and the fractions vary in different datasets. In the examination of bias, we focus on the coverage evenness rather than the populations of crystal systems. Second, $h(Y|CS) \leq h(Y)$, and equality is reached (i.e., active learning is concluded) if and only if $Y$ and $CS$ are independent. Intuitively, with $Y$ denoting $\Delta E$, this means symmetry does not provide information for stability. This is not true according to theories discussed in the introduction of structure–stability bias. Third, $h(Y|CS)$ as a single scalar value does not give a sense of the bias level, whereas the difference in $h(Y)$ of crystal systems clearly indicates how biased a dataset is.

### Information entropy estimation

For a continuous random variable $y$, evaluation of its information entropy requires the probability density function (*pdf*) $f(y)$. However, given a discrete set of values $\{y_i\}_{i=1}^n$, the underlying *pdf* is not obtainable. We use the numerical estimations implemented in scipy.stats, with the "auto" setting. Specifically, for $11 < n \leq 1000$, the result is given by the $H_e(m, n)$ estimator presented by Ebrahimi et al.[4]; while for $n > 1000$, the Vasicek estimator[5]

$$h(Y) \cong \frac{1}{n} \sum_{i=1}^{n} \log \left( \frac{n}{2m} (y_{i+m} - y_{i-m}) \right)$$

is used, where $m$, the window size, is defined by $\lfloor \sqrt{n} + 0.5 \rfloor$ in both cases.

## 2. Data preparation

### Graph embedding

The crystal graph convolutional neural network (CGCNN)[6] predicts materials properties from the graph representation of the crystal structures. The input includes node feature vector(s) that encodes atomic properties, and edge feature vector(s) that encodes connections between atoms, both can be obtained from crystallographic information framework (CIF) files without knowing other properties.

We retrieve a CGCNN model pretrained to predict the formation energy per atom ($\Delta E$) on the Materials Project dataset[7]. We feed the CIF files of materials in the J-CFID to the pretrained model and obtain the activations of the last but one layer of neurons. These 32-dimensional vectors (graph embeddings) are used as representations of J-CFID materials structures. Note that the graph embeddings do not have direct physical meanings, but they are generally obtainable for any given crystal structure.

## Physical descriptors

In supervised machine learning (ML) of mechanical properties, we use physically meaningful descriptors of materials as input. The descriptors include ones defined in the Magpie ML framework[8], the Ewald energy per atom, and volume per site. The Magpie descriptors set include the minimum, maximum, range, mean, average deviation, and mode of features such as Mendeleev number, atomic weight, and covalent radius of elements/atoms in a compound. A complete list can be found in the data files open-sourced along with the code.

The Ewald energy per atom is obtained using the "analysis.ewald" module of the pymatgen package[9]; the other descriptors are obtained using the "featurizers" module of the Matminer package[10].

## J-CFID data splitting

In preparation of the data for experiments, we first split the J-CFID dataset of size 10,898 to a test set of 4,898 datapoints and an "experiment" set of 6,000 datapoints. Due to the large size and randomness in dataset splitting, the test set has similar level of bias as the original J-CFID dataset. From the experiment set, we take (1) all tetragonal and trigonal materials with $\Delta E > 0$ and (2) all orthorhombic materials with $\Delta E < 0$, then randomly draw datapoints from the remaining data, to form an unlabeled set of size 5,000. The other 1,000 datapoints make the labeled set.

# Supplemental References

1. Pitoura, E. (2020). Social-minded Measures of Data Quality. Journal of Data and Information Quality *12*, 1-8. 10.1145/3404193.
2. Spellerberg, I.F., and Fedor, P.J. (2003). A tribute to Claude Shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'Shannon–Wiener' Index. Global Ecology and Biogeography *12*, 177-179. 10.1046/j.1466-822X.2003.00015.x.
3. Kulesza, A., and Taskar, B. (2012). Determinantal Point Processes for Machine Learning. Foundations and Trends® in Machine Learning *5*, 123-286. 10.1561/2200000044.
4. Ebrahimi, N., Pflughoeft, K., and Soofi, E.S. (1994). Two measures of sample entropy. Statistics & Probability Letters *20*, 225-234. 10.1016/0167-7152(94)90046-9.
5. Vasicek, O. (1976). A Test for Normality Based on Sample Entropy. Journal of the Royal Statistical Society. Series B (Methodological) *38*, 54-59.
6. Xie, T., and Grossman, J.C. (2018). Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. Physical Review Letters *120*, 145301. 10.1103/PhysRevLett.120.145301.
7. Jain, A., Ong, S.P., Hautier, G., Chen, W., Richards, W.D., Dacek, S., Cholia, S., Gunter, D., Skinner, D., Ceder, G., and Persson, K.A. (2013). Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL Materials *1*, 011002. 10.1063/1.4812323.
8. Ward, L., Agrawal, A., Choudhary, A., and Wolverton, C. (2016). A general-purpose machine learning framework for predicting properties of inorganic materials. npj Computational Materials *2*, 16028. 10.1038/npjcompumats.2016.28.
9. Ong, S.P., Richards, W.D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V.L., Persson, K.A., and Ceder, G. (2013). Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. Computational Materials Science *68*, 314-319. 10.1016/j.commatsci.2012.10.028.
10. Ward, L., Dunn, A., Faghaninia, A., Zimmermann, N.E.R., Bajaj, S., Wang, Q., Montoya, J., Chen, J., Bystrom, K., Dylla, M., et al. (2018). Matminer: An open source toolkit for materials data mining. Computational Materials Science *152*, 60-69. 10.1016/j.commatsci.2018.05.018.