FUSION: Dataset Pruning via Fusing Uncertainty with Structural Information for Optimal Neural Training in Crystal Property Prediction

Xiean Wang[1*], Pin Chen[1,2*], Liqin Tan[1], Yutong Lu[1,2], Qingsong Zou[1,3†]

[1]School of Computer Science and Engineering, Sun Yat-sen University, Guangzhou, China
[2]National Supercomputer Center in Guangzhou, China
[3]Guangdong Province Key Laboratory of Computational Science, Sun Yat-sen University, Guangzhou, China
{wangxan, tanlq8}@mail2.sysu.edu.cn
{chenp85, luyutong, mcszqs}@mail.sysu.edu.cn

Abstract

The rapid expansion of materials databases offers unprecedented opportunities for accelerating materials discovery via machine learning. However, the widespread assumption that larger datasets inherently produce better models does not hold in practice. We propose FUSION (**F**using **U**ncertainty with **S**tructural **I**nformation for **O**ptimal **N**eural training), an offline dataset pruning strategy that synergistically combines uncertainty quantification with crystallographic structure analysis via geometric fingerprinting, framing dataset pruning as a discrete optimization problem. Through evaluation across 3 benchmark datasets, FUSION consistently outperforms baselines, including random pruning, uncertainty sampling, weighting factor pruning, diversity sampling, and active learning. It demonstrates robust transferability across 11 diverse architectures, outperforming random pruning by 1.91–13.65% across different datasets, with an average improvement of 6.36%. Moreover, our analysis suggests that different models exhibit varying robustness characteristics when faced with pruned training data, highlighting the importance of model selection tailored to dataset composition. We identify optimal pruning points where removing just 0–8% of training data improves model performance, yielding gains up to 12.67% in specific model–dataset combinations. These results establish a new paradigm for materials informatics that prioritizes data quality over quantity, offering a pathway toward more efficient and sustainable machine learning workflows in computational materials science.

1 Introduction

The advent of high-throughput computational materials discovery has catalyzed an unprecedented expansion of materials databases, with repositories such as the Materials Project (Jain et al. 2013), Open Quantum Materials Database (Saal et al. 2013), and AFLOW (Curtarolo et al. 2012) now containing millions of crystal structures and their associated properties. This data abundance has enabled the development of increasingly sophisticated machine learning models for materials property prediction (Schmidt et al. 2019; Butler et al. 2018), yet a fundamental assumption underlying current approaches, that larger datasets inherently yield superior model performance, remains largely unexamined and potentially counterproductive.

Recent advances in materials informatics have predominantly focused on architectural innovations (Chen et al. 2024; Yan et al. 2024; Ito et al. 2025) and feature engineering (Damewood et al. 2023). In contrast, the equally important questions of optimal dataset composition and data-efficient learning paradigms, despite their growing importance amid the exponential expansion of materials databases, have received comparatively limited attention (Li et al. 2024; Lookman et al. 2019; Li et al. 2023; Wang, Chen, and Zou 2025). The fundamental challenge in pruning datasets for materials property prediction lies in the complex interplay of structure, uncertainty, and learning behavior. Effective pruning must account for (1) predictive uncertainty, which reflects both model confidence and data quality (Gal and Ghahramani 2016; Lakshminarayanan, Pritzel, and Blundell 2017); (2) crystallographic and chemical diversity essential for covering materials space (Meredig et al. 2018; Janet et al. 2019); and (3) the architecture-specific learning dynamics inherent to different neural networks (Ramprasad et al. 2017; Schmidt et al. 2019).

To address these limitations, we propose FUSION (**F**using **U**ncertainty with **S**tructural **I**nformation for **O**ptimal **N**eural training), an offline dataset pruning methodology that synergistically combines uncertainty quantification with crystallographic structure analysis. FUSION formulates the dataset pruning as a discrete optimization problem that balances predictive uncertainty with structural representativeness, enabling identification of redundant samples while preserving essential learning information. In our work, we also establish a data preprocessing framework compatible with model training, which can be integrated with any existing machine learning models for material property prediction.

Our main contributions are summarized as follows:

- We propose FUSION. To the best of our knowledge, this is the first material dataset pruning method that simultaneously considers model cognition and essential material structure information.
- Through comprehensive evaluation across 3 benchmark

*These authors contributed equally.
†Corresponding author.

datasets and 11 model architectures, we systematically discover and quantify the "optimal pruning points" phenomenon for the first time in materials science.

- We develop a complete end-to-end training pipeline that seamlessly integrates dataset pruning into the existing model training workflow, providing a systematic approach for materials property prediction.

2 Related Work

2.1 Data Redundancy

Materials databases accumulate near-equivalent structures, such as polymorphic variations, minor lattice distortions, and systematically enumerated structural decorations, which make up a substantial fraction of the database content (Haastrup et al. 2018; Zhou et al. 2019). This redundancy manifests particularly in databases like Materials Project (Jain et al. 2013), where individual compositions such as Al_2O_3 exhibit 15 distinct crystallographic representations spanning a range of thermodynamic stabilities (Oguchi 2024). Materials databases exhibit unique characteristics where chemically similar compounds can display vastly different electronic and thermodynamic properties, as demonstrated by high-throughput DFT screening studies conducted by Saal et al. (2013). Conventional structural similarity-based pruning methods fail to distinguish between geometrically alike structures that provide genuine learning value versus those that introduce noise, leading to models that overfit to specific atomic arrangements rather than generalizable structure-property relationships (Meredig et al. 2018). This necessitates pruning strategies that simultaneously consider both crystallographic similarity and predictive learning dynamics to optimize dataset composition for robust materials property prediction.

2.2 Similarity and Uncertainty

Geometric descriptors can be used to measure the similarity of crystal structures. They have evolved from simple structural features to sophisticated rotationally invariant representations. Early approaches relied on neural network representations of potential energy surfaces (Behler and Parrinello 2007) and handcrafted molecular descriptors such as Coulomb matrices (Rupp et al. 2012), which provided intuitive but limited representations of atomic interactions. A significant advancement came with the introduction of Smooth Overlap of Atomic Positions (SOAP) descriptors (Bartók, Kondor, and Csányi 2013), which provided smooth, differentiable representations of local atomic environments that respect rotational and translational symmetry. Building upon these foundations, subsequent developments have extended geometric representations to include more sophisticated approaches: Ward et al. (2016) developed a general-purpose machine learning framework incorporating chemically diverse attributes for materials property prediction, while Pham et al. (2017) introduced orbital interaction-based descriptors such as the Orbital Field Matrix (OFM). More recent work has focused on incorporating global structural information through techniques such as Crystal Graph Convolutional Networks (Xie and Grossman 2018) and advanced many-body representations (Huo and Rupp 2022). The DScribe library (Himanen et al. 2020) has standardized many of these approaches, enabling systematic comparisons and reproducible research across different descriptor methodologies. When applied to dataset pruning, typical practices such as k-means clustering or diversity sampling based on geometric descriptors aim to ensure structural diversity (Abbas et al. 2024). However, these approaches inherently assume that geometric diversity correlates with learning value, an assumption that may not hold under the complex dynamics of neural network training (Zheng et al. 2023).

Uncertainty quantification (UQ) has become a crucial aspect of reliable materials property prediction, particularly in scenarios where prediction confidence directly influences experimental decision-making (Lookman et al. 2019; Raccuglia et al. 2016). Traditional UQ approaches, such as deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017) and Monte Carlo dropout (Gal and Ghahramani 2016), provide uncertainty estimates but often come with substantial computational overhead. Deep evidential regression (DER) (Amini et al. 2020; Sensoy, Kaplan, and Kandemir 2018) offers a promising alternative by estimating both aleatoric (data-driven) and epistemic (model-driven) uncertainty in a single forward pass. Its applications in materials science have demonstrated enhanced calibration and improved out-of-distribution detection (Soleimany et al. 2021). In light of the above, UQ in materials machine learning has been primarily applied to improve prediction accuracy and model reliability, with limited exploration into using uncertainty estimates to guide dataset pruning.

2.3 Dataset Pruning

Dataset pruning has received growing attention across machine learning domains such as computer vision (Paul, Ganguli, and Dziugaite 2021; Sorscher et al. 2022) and NLP (Abbas et al. 2023; Xia et al. 2024). The limited body of work addressing materials dataset pruning has primarily focused on coarse-grained approaches that sacrifice precision for computational efficiency. Li et al. (2023) investigated redundancy in large-scale materials datasets using uncertainty measures derived from model ensemble disagreement or simply the mean absolute error (MAE) value of a single model's predicted attribute, demonstrating that significant dataset reductions are possible without substantial performance loss. Li et al. (2024) proposed MD-HIT, a greedy redundancy reduction algorithm inspired by bioinformatics clustering techniques, utilizing composition-based and structure-based distance metrics to systematically eliminate redundant samples above predefined similarity thresholds. However, their approaches rely on relatively simple metrics and do not account for the complex interplay between structural similarity and learning value.

Although structural analysis tools and uncertainty estimation techniques have both advanced significantly in materials research, they remain largely siloed in practice. Crystallographic descriptors provide physically grounded representations but are seldom aligned with the data effectiveness

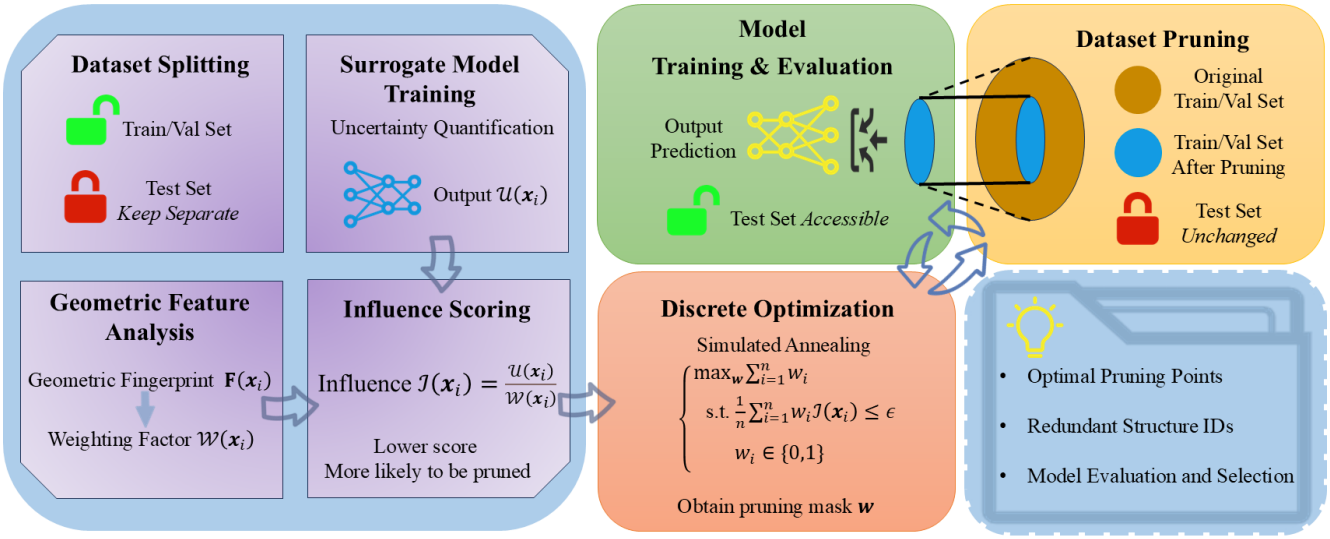

Figure 1: Framework and training pipeline overview of the proposed FUSION dataset pruning method.

goals of neural network training. Conversely, uncertainty-aware methods improve model calibration, yet rarely inform decisions about which training data are worth keeping. Few existing approaches combine these perspectives into a cohesive pruning strategy. In this work, we bridge this divide by jointly leveraging predictive uncertainty and structural similarity to perform fine-grained, model-agnostic dataset pruning, enabling more efficient and principled learning in materials informatics.

3 Method

3.1 Problem Formulation

We define an ϵ-redundant subset $\hat{\mathcal{D}}_\epsilon \subseteq \mathcal{D}$ as a subset whose removal results in model generalization performance degradation of at most ϵ:

$$\mathcal{L}(\hat{\boldsymbol{\theta}}_{-\hat{\mathcal{D}}_\epsilon}) - \mathcal{L}(\hat{\boldsymbol{\theta}}) \leq \epsilon \tag{1}$$

where $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_{-\hat{\mathcal{D}}_\epsilon}$ are the optimal model parameters trained on the full dataset $\mathcal{D}$ and the pruned dataset $\mathcal{D} \setminus \hat{\mathcal{D}}_\epsilon$, respectively, and $\mathcal{L}(\cdot)$ denotes the expected loss over the true data distribution. In practice, $\mathcal{L}(\cdot)$ is estimated using validation or test set performance.

The dataset pruning problem then becomes finding the largest ϵ-redundant subset:

$$\hat{\mathcal{D}}_\epsilon^{\max} = \arg\max_{\hat{\mathcal{D}}_\epsilon \subseteq \mathcal{D}} |\hat{\mathcal{D}}_\epsilon| \quad \text{s.t.} \quad \mathcal{L}(\hat{\boldsymbol{\theta}}_{-\hat{\mathcal{D}}_\epsilon}) - \mathcal{L}(\hat{\boldsymbol{\theta}}) \leq \epsilon \tag{2}$$

3.2 Uncertainty Estimation

Our approach uses uncertainty to quantify the influence of each crystal structure on model generalization. The data-based uncertainty $\mathcal{U}_{\text{ale}}(\boldsymbol{x}_i)$ and the model-based uncertainty $\mathcal{U}_{\text{epi}}(\boldsymbol{x}_i)$ are estimated using DER (Amini et al. 2020). For a given crystal structure $\boldsymbol{x}_i$, the model outputs parameters of a Normal-Inverse-Gamma (NIG) distribution, and the total uncertainty is computed as:

$$\begin{aligned}\mathcal{U}(\boldsymbol{x}_i) &= \text{sigmoid}(-\mathcal{U}_{\text{ale}}(\boldsymbol{x}_i) + \lambda \mathcal{U}_{\text{epi}}(\boldsymbol{x}_i)) \\ &= \text{sigmoid}\left(-\frac{\beta_i}{\alpha_i - 1} + \lambda \cdot \frac{\beta_i}{(\alpha_i - 1)\nu_i}\right)\end{aligned} \tag{3}$$

where α_i, β_i, and ν_i are the NIG distribution parameters output by the model for input structure $\boldsymbol{x}_i$. The sigmoid function ensures the uncertainty estimate is bounded in $(0, 1)$, enabling the estimated uncertainty to remain strictly positive and providing numerical stability. Effective dataset pruning requires distinguishing between aleatoric and epistemic uncertainties rather than simply removing high-uncertainty samples. The parameter λ in Equation 3 controls this distinction: when $\lambda \to 0$, the formulation emphasizes aleatoric uncertainty to eliminate data noise, while larger λ values emphasize epistemic uncertainty to reduce model knowledge gaps.

3.3 Geometric Feature-Based Weighting Factor

For each crystal structure, we compute a geometric feature fingerprint $\mathbf{F}(\boldsymbol{x}_i)$ that captures both local atomic environments and global structural characteristics. This fingerprint is derived using the SOAP descriptor (Bartók, Kondor, and Csányi 2013), which provides rotationally and translationally invariant representations of crystal structures.

The computation of $\mathbf{F}(\boldsymbol{x}_i)$ involves several key steps. First, we identify equivalent atomic sites within the crystal structure using space group symmetry analysis. For a crystal structure, atoms that are related by symmetry operations occupy equivalent sites and contribute identically to the material's properties. We group atoms into equivalent site classes $\{S_{i1}, S_{i2}, \ldots, S_{ik_i}\}$, where each class S_{ij} contains m_{ij} symmetrically equivalent atoms. Next, we compute SOAP descriptors $\mathbf{s}_{ij}$ for each equivalent site group.

The geometric feature fingerprint is then computed as a weighted average over all equivalent site groups:

$$\mathbf{F}(\boldsymbol{x}_i) = \frac{1}{N_i} \sum_{j=1}^{k_i} m_{ij} \mathbf{s}_{ij} \tag{4}$$

where $N_i = \sum_{j=1}^{k_i} m_{ij}$ is the total number of atoms in the unit cell. Finally, we apply L2 normalization to the geometric fingerprint to ensure scale invariance.

Using the geometric feature fingerprints, we define the Weighting Factor as:

$$\mathcal{W}(\boldsymbol{x}_i) = \exp\left(-\gamma \cdot \min_{j \neq i} d(\boldsymbol{x}_i, \boldsymbol{x}_j)\right) \tag{5}$$

where $d(\boldsymbol{x}_i, \boldsymbol{x}_j) = 1 - \mathbf{F}(\boldsymbol{x}_i)^T \mathbf{F}(\boldsymbol{x}_j)$ is the cosine distance between geometric fingerprints, and γ is a hyperparameter controlling the influence of structural similarity. This Weighting Factor increases for structures that are similar to others in the dataset, penalizing redundant structural information, while decreasing for structurally isolated examples that may represent unique material classes.

3.4 Optimization-Based Dataset Pruning

We define the influence of a structure $\boldsymbol{x}_i$ as:

$$\mathcal{I}(\boldsymbol{x}_i) = \frac{\mathcal{U}(\boldsymbol{x}_i)}{\mathcal{W}(\boldsymbol{x}_i)} \tag{6}$$

where $\mathcal{U}(\boldsymbol{x}_i)$ is the predictive uncertainty for structure $\boldsymbol{x}_i$ and $\mathcal{W}(\boldsymbol{x}_i)$ is the Weighting Factor that accounts for the structure's representativeness in the dataset.

Directly solving this optimization problem described in Equation 2 is computationally intractable as it requires enumerating $2^{|\mathcal{D}|}$ possible subsets and retraining the model for each subset evaluation. The key insight of our approach is to replace the expensive retraining requirement with influence scores $\mathcal{I}(\boldsymbol{x}_i)$, enabling scalable approximation of the original NP-hard problem. We formulate the dataset pruning problem as a discrete optimization that maximizes the number of removed structures while constraining the collective impact on model performance:

$$\begin{aligned} \max_{\boldsymbol{w}} \quad & \sum_{i=1}^{n} w_i \\ \text{s.t.} \quad & \frac{1}{n} \sum_{i=1}^{n} w_i \mathcal{I}(\boldsymbol{x}_i) \leq \epsilon \\ & w_i \in \{0, 1\} \end{aligned} \tag{7}$$

where $w_i = 1$ indicates that structure $\boldsymbol{x}_i$ is selected for removal, and $w_i = 0$ indicates retention. The constraint ensures that the collective influence of removed structures remains bounded by ϵ, limiting the performance degradation. To solve this discrete optimization problem, we employ a simulated annealing algorithm (Kirkpatrick, Gelatt Jr, and Vecchi 1983).

3.5 Iterative Optimization for Optimal Training

Figure 1 presents the FUSION framework, which comprises 7 stages organized within an iterative optimization loop. The process begins with Dataset Splitting, where data is partitioned into training, validation, and test sets, with the test set strictly held out. In the second stage, Surrogate Model Training, a DER model is fitted to estimate predictive uncertainty. Next, Geometric Feature Analysis extracts structural fingerprints using SOAP descriptors combined with symmetry-aware aggregation. In the fourth stage, Influence Scoring, uncertainty estimates are integrated with similarity-based weighting to assess each sample's contribution to generalization performance.

The final three stages, Discrete Optimization, Dataset Pruning, and Model Training and Evaluation, form a closed-loop system that iteratively explores pruning configurations by adjusting the tolerance parameter ϵ. Through this cyclic process, FUSION identifies the Optimal Pruning Point at which targeted dataset pruning leads to peak model performance. This enables comprehensive evaluation across multiple architectures and facilitates model selection tailored to varying dataset compositions.

4 Experiments

4.1 Experimental Setup

We evaluate our approach on three materials property prediction tasks from the well-established Matbench benchmark (Dunn et al. 2020): **Dielectric**, with 4,764 samples, for predicting refractive index from crystal structure; **JDFT2D**, with 636 samples, for predicting exfoliation energies of 2D layered materials; and **Perovskites**, with 18,928 samples, for predicting formation energies of perovskite structures. These datasets cover diverse material systems and property types, including optical, mechanical, and thermodynamic properties, providing a representative benchmark for evaluating model generalization and robustness.

All experiments, except for the out-of-distribution (OOD) test, use fixed dataset splits of 70% for training, 15% for validation, and 15% for testing. To ensure fair comparison, the test set remains identical across all pruning methods and ratios in the same task, while training and validation sets undergo different pruning strategies. Each configuration is evaluated with five random seeds to enable statistical significance testing. We use ALIGNN (Choudhary and DeCost 2021) combined with DER (Amini et al. 2020) as our surrogate model for generating uncertainty. The pruned datasets are then used to train and test ALIGNN. In the cross-model experiments, we set $\lambda = 0$ to eliminate the impact of model-based uncertainty $\mathcal{U}_{\text{epi}}(\boldsymbol{x}_i)$, and only retain data-based uncertainty $\mathcal{U}_{\text{ale}}(\boldsymbol{x}_i)$. Then we use the surrogate model ALIGNN to prune the dataset for training all models of different architectures. To prevent data leakage and ensure a fair evaluation, the surrogate model was trained without any access to the test set in all experiments.

We assess our pruning methods across 11 architectures for materials property prediction. The models encompass graph-based approaches, including CGCNN (Xie and Grossman 2018), SchNet (Schütt et al. 2018), MEGNet

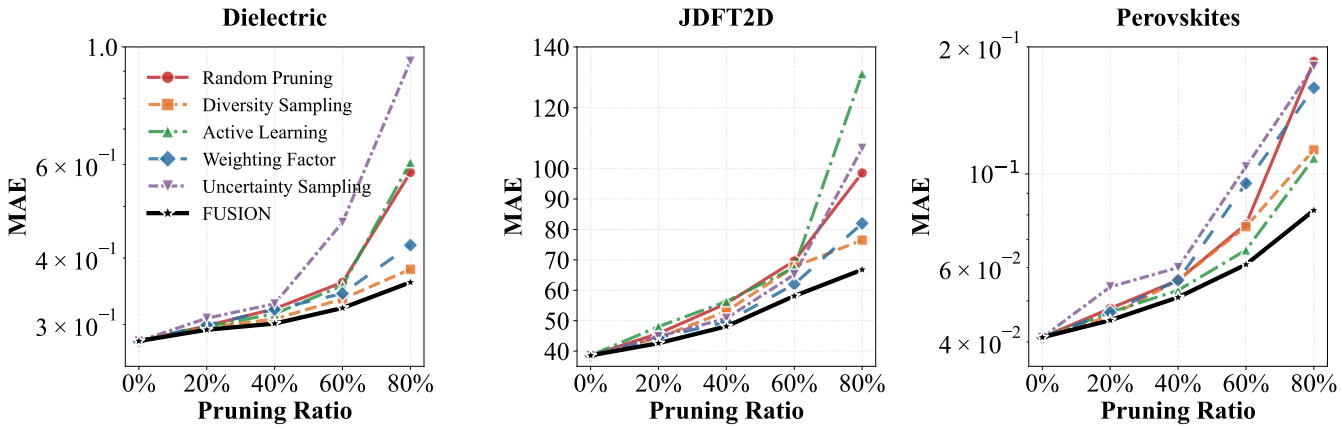

Figure 2: In-distribution MAE comparison across pruning methods on three Matbench datasets. **(Left)** Dielectric (unitless); **(Middle)** JDFT2D (meV/atom); **(Right)** Perovskites (eV). Lines show the mean MAE over five random seeds.

(Chen et al. 2019), ALIGNN (Choudhary and DeCost 2021), and PotNet (Lin et al. 2023), as well as transformer-based architectures such as DeeperGATGNN (Omee et al. 2022), Matformer (Yan et al. 2022), eComFormer (Yan et al. 2024), iComFormer (Yan et al. 2024), SODNet (Chen et al. 2024), and CrystalFramer (Ito et al. 2025). This diversity ensures the generalizability of our pruning methodology across different neural network architectures. To ensure fair comparison, identical test sets are used across all models.

We construct OOD splits using leave-one-cluster-out cross-validation (Meredig et al. 2018). Materials are first grouped into clusters via k-means clustering on SOAP descriptors, then each cluster is iteratively held out as the test set while training on the remaining clusters. We select cluster partitions that yield the most challenging evaluation scenarios: 40 clusters for Dielectric and JDFT2D datasets, and 5 clusters for Perovskites. We report the averaged MAE across the first 5 iterations in the following experimental sections.

We compare FUSION against five baseline methods. **Random Pruning** serves as the fundamental baseline, implementing purely stochastic structure selection without domain knowledge. **Diversity Sampling** utilizes Farthest Point Sampling (FPS) (Eldar et al. 1997; Sener and Savarese 2017) on SOAP descriptors to maintain structural diversity, prioritizing coverage of the materials space through geometric distance maximization. **Active Learning** implements Query-by-Committee using an ensemble (Beluch et al. 2018) of five ALIGNN models, selecting structures based on prediction disagreement to identify the most informative examples. Finally, **Weighting Factor** isolates the structural similarity component of FUSION by implementing pruning based solely on geometric relationships, without incorporating uncertainty information, thereby enabling an independent assessment of the contribution from structural analysis. In contrast, **Uncertainty Sampling** serves as a complementary baseline that relies exclusively on uncertainty estimates. It employs DER (Amini et al. 2020) to compute predictive uncertainty and prunes data points with the lowest uncertainty values using Equation 3.

4.2 Experimental Results

Performance on Benchmark Datasets As illustrated in Figure 2, FUSION consistently achieves the lowest MAE across all pruning ratios. The performance advantages become increasingly pronounced at higher pruning ratios. For the Dielectric dataset, FUSION demonstrates progressive performance gains relative to the second-best method, Diversity Sampling, with improvements ranging from 1.0% at the 20% pruning ratio to 5.5% at the 80% pruning ratio. The JDFT2D dataset exhibits more substantial performance differences, with FUSION achieving a remarkable 12.7% MAE reduction compared to diversity sampling at 80% pruning.

Diversity Sampling consistently achieves competitive performance across most scenarios, validating the importance of maintaining structural coverage in the materials space. However, its exclusive focus on geometric diversity without consideration of prediction difficulty becomes apparent as a limitation in challenging scenarios where complex structure-property relationships require more nuanced data selection strategies. Active Learning shows considerable instability, particularly evident in the JDFT2D dataset, where the MAE jumps to 131.210 at 80% pruning. This instability arises from the ensemble disagreement metric's sensitivity to individual model variations and training stochasticity, resulting in inconsistent assessments of sample importance. The disagreement-based selection criterion, while theoretically sound for identifying informative samples in traditional active learning contexts, proves less reliable for dataset pruning applications where the objective is to identify and remove redundant rather than informative samples.

Generalization Under Distribution Shift Out-of-distribution (OOD) evaluation aims to simulate realistic materials discovery scenarios where models must generalize across different chemical spaces. The results in Table 1 reveal that FUSION maintains superior performance across all configurations, though the relative advantages among different methods show notable changes compared to the in-distribution (ID) setting. We conducted paired t-tests over 5 runs for all ID and OOD experiments. While FUSION

achieved statistically significant improvements over the second-best baseline in the majority of cases with $p < 0.05$ in the OOD setting, these differences were generally weaker than those observed in the ID setting, where p-values were typically below 0.01. Notably, at a high pruning ratio of 80%, the ID results remained highly statistically significant, with p-values below 0.001. This contrast highlights the inherent difficulty of achieving strong generalization in the OOD setting.

The OOD results reveal a convergence in performance across methods compared to the ID scenario. For instance, in the JDFT2D dataset at 80% pruning, the performance gap between FUSION and the second-best baseline reduces from 9.752 MAE in the ID setting to only 0.185 MAE in the OOD setting. This convergence pattern occurs consistently across datasets and suggests that distribution shift introduces fundamental challenges that reduce the advantages of sophisticated data selection strategies. The phenomenon indicates that when models encounter novel chemical spaces, the benefits of careful training dataset pruning become less pronounced, as all methods struggle with the inherent difficulty of extrapolating beyond their training distributions.

Dataset	Method	w/o pruning	20%	40%	60%	80%
Dielectric	Random Pruning	**0.247**	0.267	0.287	0.367	0.421
	Diversity Sampling		0.268	0.269	0.302	<u>0.368</u>
	Active Learning		<u>0.265</u>	0.274	<u>0.301</u>	0.522
	Weighting Factor		0.266	<u>0.268</u>	0.303	0.373
	Uncertainty Sampling		0.266	0.297	0.490	0.791
	FUSION		**0.264**	**0.265**	**0.274**	**0.319**
JDFT2D	Random Pruning	**34.115**	54.646	61.806	67.028	96.598
	Diversity Sampling		43.749	51.029	53.747	79.947
	Active Learning		47.295	53.177	66.590	77.425
	Weighting Factor		51.723	63.389	65.905	<u>69.470</u>
	Uncertainty Sampling		<u>42.190</u>	<u>45.116</u>	<u>50.044</u>	72.903
	FUSION		**38.538**	**43.771**	**49.164**	**69.285**
Perovskites	Random Pruning	**0.077**	0.086	0.098	0.200	0.223
	Diversity Sampling		0.084	0.093	0.119	0.155
	Active Learning		0.085	<u>0.092</u>	<u>0.104</u>	<u>0.149</u>
	Weighting Factor		<u>0.084</u>	0.097	0.142	0.200
	Uncertainty Sampling		0.113	0.148	0.197	0.293
	FUSION		**0.083**	**0.090**	**0.100**	**0.127**

Table 1: MAE comparison in OOD test across different datasets, pruning methods, and dataset pruning ratios. Best results for each dataset-ratio combination are highlighted in **bold**, and second-best results are <u>underlined</u>.

Ablation Study Uncertainty Sampling represents the isolated use of the DER uncertainty prediction component from FUSION. However, as illustrated by Table 1 and Figure 2, Uncertainty Sampling exhibits significant performance degradation at high pruning ratios in both ID and OOD tests, most notably in the ID test of the Dielectric dataset, where the MAE increases to 0.940 at 80% pruning, representing a 161% deterioration compared to FUSION. This phenomenon exposes the inherent weakness of relying solely on prediction uncertainty for dataset pruning, as removing low-uncertainty samples can inadvertently eliminate crucial structural information that supports model training stability. The samples with low predictive uncertainty often represent well-understood regions of the materials space, and their removal can paradoxically reduce the model's ability to generalize to related but distinct structural configurations. The method's reliance on model confidence measures that are calibrated on the training distribution is unreliable when applied to OOD samples. This failure mode highlights a critical limitation of uncertainty-based pruning strategies and validates FUSION's incorporation of distribution-agnostic structural information as a crucial robustness mechanism. The Weighting Factor baseline achieves competitive performance at moderate pruning ratios but experiences significant degradation at aggressive pruning levels. While structural similarity provides valuable guidance for identifying redundant crystal structures, it lacks the predictive awareness necessary to distinguish between geometrically similar structures that may exhibit vastly different learning contributions.

Cross-Model Consistency and Transferability We train 11 different models on datasets pruned by FUSION and Random Pruning to evaluate the consistency and transferability of our pruning strategy across diverse neural network architectures. Table 2 presents a comparison of performance improvements achieved by FUSION relative to Random Pruning, averaged across all 11 model architectures.

Dataset	20%	40%	60%	80%	Average
Dielectric	1.66%	1.38%	1.37%	3.24%	1.91%
JDFT2D	3.27%	3.06%	4.97%	2.76%	3.52%
Perovskites	15.78%	15.98%	12.75%	10.08%	13.65%
Average	6.91%	6.81%	6.36%	5.36%	6.36%

Table 2: Detailed performance improvement analysis. Values represent the average improvement percentage across all 11 models. Positive values indicate FUSION outperforms Random Pruning.

The results demonstrate consistency in FUSION's superior performance across diverse model architectures, with average improvements ranging from 1.91% to 13.65% depending on the dataset characteristics. The Perovskites dataset exhibits the most substantial improvements, achieving an average enhancement of 13.65% across all pruning ratios, which can be attributed to the dataset's high redundancy and the effectiveness of our approach in identifying redundant crystal structures. The JDFT2D dataset shows moderate but consistent improvements averaging 3.52%, while the Dielectric dataset demonstrates smaller improvements of 1.91%, suggesting that different material property prediction tasks exhibit varying levels of dataset redundancy and thus different potential for improvement through strategic pruning. The performance advantages tend to be most pronounced at moderate pruning ratios (20%-60%), with some diminishment at aggressive 80% pruning.

The cross-model evaluation reveals a phenomenon where relative model rankings exhibit sensitivity to the specific

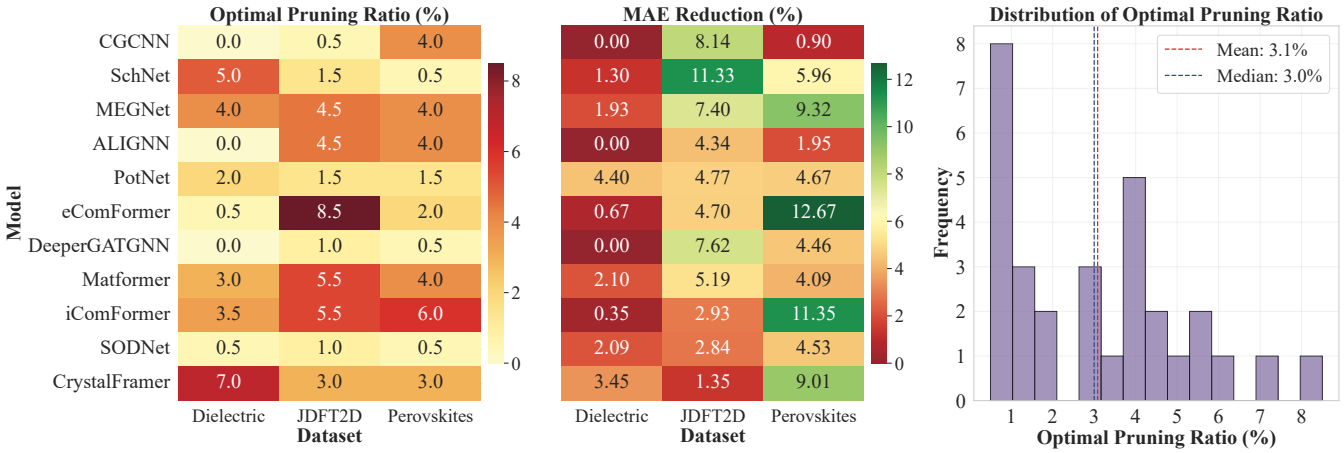

Figure 3: Analysis of optimal pruning ratio across models and datasets. We incrementally remove data in 0.5% steps to identify optimal pruning ratios. **(Left)** Optimal pruning ratios for each model-dataset pair. **(Middle)** MAE reduction percentages achieved at optimal ratios relative to unpruned baselines. **(Right)** Distribution of optimal pruning ratios, with pruning ratios of 0 excluded from the analysis.

data subset used for training, demonstrating that optimal model selection is inherently dependent on the dataset composition rather than representing universal hierarchies. This ranking variability manifests clearly across our experiments: when training on data pruned using FUSION, in the Dielectric dataset, ALIGNN achieves superior performance at pruning ratios of 20%, 40%, 60%, and 80%, while iComFormer demonstrates the best performance on the complete unpruned dataset. Similar ranking shifts occur in the JDFT2D dataset, where iComFormer dominates most pruning scenarios except at 80% pruning, where PotNet takes the lead. This phenomenon suggests that different models exhibit varying robustness characteristics when faced with reduced training data, with some architectures maintaining performance under data scarcity while others excel only with abundant training examples. The observed ranking volatility underscores a fundamental limitation of model evaluation practices that rely on single dataset configurations, highlighting the importance of data-aware model selection strategies that consider the specific characteristics and size of available training data.

Optimal Pruning Ratio Analysis The relationship between dataset size and model performance is not strictly monotonic. As illustrated in Figure 3, the optimal pruning ratios exhibit remarkable diversity across model architectures and datasets, with a mean of 3.1% and a median of 3.0%, yet spanning a range from 0% to 8.5%. This distribution suggests the existence of distinct optimal pruning points where strategic dataset pruning enhances model performance, with notable examples including eComFormer achieving 12.67% MAE reduction on the Perovskites dataset at 2.0% pruning and SchNet demonstrating 11.33% improvement on JDFT2D at 1.5% pruning relative to the unpruned datasets. The marginal utility of additional training examples follows a diminishing returns pattern that can become negative when low-quality or redundant structures dominate the dataset. The heterogeneity in optimal pruning ratios across architectures reflects each model's distinct sensitivity to data quality versus quantity tradeoffs, where transformer-based models like eComFormer and iComFormer demonstrate greater resilience to aggressive pruning compared to traditional graph neural networks.

The existence of the optimal pruning points can be attributed to the intrinsic noise and redundancy present in large-scale materials databases, where similar crystal structures with minor variations may introduce conflicting learning signals that impede model convergence and generalization. FUSION can serve not merely as a dataset pruning technique but as a data quality enhancement methodology that transforms the fundamental relationship between dataset pruning and model effectiveness in computational materials science.

5 Conclusion

This paper proposes FUSION, a dataset pruning strategy that addresses dataset redundancy in crystal property prediction by strategically removing training samples with minimal impact on model performance. Our work simultaneously integrates model cognition via uncertainty quantification with essential crystallographic structure information through geometric fingerprinting. We systematically discover and quantify the optimal pruning points in materials science, demonstrating that strategic dataset pruning can enhance model performance. Additionally, we develop a comprehensive end-to-end training pipeline that seamlessly incorporates dataset pruning into existing materials property prediction workflows. This paper establishes a new paradigm for materials informatics that emphasizes intelligent data curation over mere data accumulation and provides a foundation for more efficient computational materials discovery.

Acknowledgments

This work was supported by the National Science and Technology Major Project (2022ZD0117805), by the National Natural Science Foundation of China under grants 92370113, and by the Guangdong Provincial Key Laboratory of Mathematical Foundations for Artificial Intelligence (2023B1212010001).

References

Abbas, A.; Rusak, E.; Tirumala, K.; Brendel, W.; Chaudhuri, K.; and Morcos, A. S. 2024. Effective pruning of web-scale datasets based on complexity of concept clusters. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*.

Abbas, A.; Tirumala, K.; Simig, D.; Ganguli, S.; and Morcos, A. S. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*.

Amini, A.; Schwarting, W.; Soleimany, A.; and Rus, D. 2020. Deep evidential regression. *Advances in neural information processing systems*, 33: 14927–14937.

Bartók, A. P.; Kondor, R.; and Csányi, G. 2013. On representing chemical environments. *Physical Review B—Condensed Matter and Materials Physics*, 87(18): 184115.

Behler, J.; and Parrinello, M. 2007. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters*, 98(14): 146401.

Beluch, W. H.; Genewein, T.; Nürnberger, A.; and Köhler, J. M. 2018. The power of ensembles for active learning in image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9368–9377.

Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; and Walsh, A. 2018. Machine learning for molecular and materials science. *Nature*, 559(7715): 547–555.

Chen, C.; Ye, W.; Zuo, Y.; Zheng, C.; and Ong, S. P. 2019. Graph networks as a universal machine learning framework for molecules and crystals. *Chemistry of Materials*, 31(9): 3564–3572.

Chen, P.; Peng, L.; Jiao, R.; Mo, Q.; Zhen, W.; Huang, W.; Liu, Y.; and Lu, Y. 2024. Learning Superconductivity from Ordered and Disordered Material Structures. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Choudhary, K.; and DeCost, B. 2021. Atomistic line graph neural network for improved materials property predictions. *npj Computational Materials*, 7(1): 185.

Curtarolo, S.; Setyawan, W.; Hart, G. L.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; et al. 2012. AFLOW: An automatic framework for high-throughput materials discovery. *Computational Materials Science*, 58: 218–226.

Damewood, J.; Karaguesian, J.; Lunger, J. R.; Tan, A. R.; Xie, M.; Peng, J.; and Gómez-Bombarelli, R. 2023. Representations of materials for machine learning. *Annual Review of Materials Research*, 53(1): 399–426.

Dunn, A.; Wang, Q.; Ganose, A.; Dopp, D.; and Jain, A. 2020. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Computational Materials*, 6(1): 138.

Eldar, Y.; Lindenbaum, M.; Porat, M.; and Zeevi, Y. Y. 1997. The farthest point strategy for progressive image sampling. *IEEE transactions on image processing*, 6(9): 1305–1315.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059. PMLR.

Haastrup, S.; Strange, M.; Pandey, M.; Deilmann, T.; Schmidt, P. S.; Hinsche, N. F.; Gjerding, M. N.; Torelli, D.; Larsen, P. M.; Riis-Jensen, A. C.; et al. 2018. The Computational 2D Materials Database: high-throughput modeling and discovery of atomically thin crystals. *2D Materials*, 5(4): 042002.

Himanen, L.; Jäger, M. O.; Morooka, E. V.; Canova, F. F.; Ranawat, Y. S.; Gao, D. Z.; Rinke, P.; and Foster, A. S. 2020. DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications*, 247: 106949.

Huo, H.; and Rupp, M. 2022. Unified representation of molecules and crystals for machine learning. *Machine Learning: Science and Technology*, 3(4): 045017.

Ito, Y.; Taniai, T.; Igarashi, R.; Ushiku, Y.; and Ono, K. 2025. Rethinking the Role of Frames for SE(3)-Invariant Crystal Structure Modeling. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*.

Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; et al. 2013. Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. *APL materials*, 1(1).

Janet, J. P.; Duan, C.; Yang, T.; Nandy, A.; and Kulik, H. J. 2019. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chemical science*, 10(34): 7913–7922.

Kirkpatrick, S.; Gelatt Jr, C. D.; and Vecchi, M. P. 1983. Optimization by simulated annealing. *science*, 220(4598): 671–680.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Li, K.; Persaud, D.; Choudhary, K.; DeCost, B.; Greenwood, M.; and Hattrick-Simpers, J. 2023. Exploiting redundancy in large materials datasets for efficient machine learning with less data. *Nature Communications*, 14(1): 7283.

Li, Q.; Fu, N.; Omee, S. S.; and Hu, J. 2024. MD-HIT: Machine learning for material property prediction with dataset redundancy control. *npj Computational Materials*, 10(1): 245.

Lin, Y.; Yan, K.; Luo, Y.; Liu, Y.; Qian, X.; and Ji, S. 2023. Efficient Approximations of Complete Interatomic Potentials for Crystal Property Prediction. In *Proceedings of the 40th International Conference on Machine Learning*.

Lookman, T.; Balachandran, P. V.; Xue, D.; and Yuan, R. 2019. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *npj Computational Materials*, 5(1): 21.

Meredig, B.; Antono, E.; Church, C.; Hutchinson, M.; Ling, J.; Paradiso, S.; Blaiszik, B.; Foster, I.; Gibbons, B.; Hattrick-Simpers, J.; et al. 2018. Can machine learning identify the next high-temperature superconductor? Examining extrapolation performance for materials discovery. *Molecular Systems Design & Engineering*, 3(5): 819–825.

Oguchi, T. 2024. Crystal structure map for materials classification and modeling. *Science and Technology of Advanced Materials: Methods*, 4(1): 2355860.

Omee, S. S.; Louis, S.-Y.; Fu, N.; Wei, L.; Dey, S.; Dong, R.; Li, Q.; and Hu, J. 2022. Scalable deeper graph neural networks for high-performance materials property prediction. *Patterns*, 3(5).

Paul, M.; Ganguli, S.; and Dziugaite, G. K. 2021. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34: 20596–20607.

Pham, T. L.; Kino, H.; Terakura, K.; Miyake, T.; Tsuda, K.; Takigawa, I.; and Dam, H. C. 2017. Machine learning reveals orbital interaction in materials. *Science and technology of advanced materials*, 18(1): 756.

Raccuglia, P.; Elbert, K. C.; Adler, P. D.; Falk, C.; Wenny, M. B.; Mollo, A.; Zeller, M.; Friedler, S. A.; Schrier, J.; and Norquist, A. J. 2016. Machine-learning-assisted materials discovery using failed experiments. *Nature*, 533(7601): 73–76.

Ramprasad, R.; Batra, R.; Pilania, G.; Mannodi-Kanakkithodi, A.; and Kim, C. 2017. Machine learning in materials informatics: recent applications and prospects. *npj Computational Materials*, 3(1): 54.

Rupp, M.; Tkatchenko, A.; Müller, K.-R.; and Von Lilienfeld, O. A. 2012. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5): 058301.

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(11): 1501–1509.

Schmidt, J.; Marques, M. R.; Botti, S.; and Marques, M. A. 2019. Recent advances and applications of machine learning in solid-state materials science. *npj computational materials*, 5(1): 83.

Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; and Müller, K.-R. 2018. Schnet–a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24).

Sener, O.; and Savarese, S. 2017. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems*, 31.

Soleimany, A. P.; Amini, A.; Goldman, S.; Rus, D.; Bhatia, S. N.; and Coley, C. W. 2021. Evidential deep learning for guided molecular property prediction and discovery. *ACS central science*, 7(8): 1356–1367.

Sorscher, B.; Geirhos, R.; Shekhar, S.; Ganguli, S.; and Morcos, A. 2022. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35: 19523–19536.

Wang, X.; Chen, P.; and Zou, Q. 2025. No-Data-Driven Crystal Structure Prediction via Model-Free Reinforcement Learning. In *International Conference on Intelligent Computing*, 307–318. Springer.

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(1): 1–7.

Xia, M.; Malladi, S.; Gururangan, S.; Arora, S.; and Chen, D. 2024. Less: Selecting influential data for targeted instruction tuning. *arXiv preprint arXiv:2402.04333*.

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(14): 145301.

Yan, K.; Fu, C.; Qian, X.; Qian, X.; and Ji, S. 2024. Complete and Efficient Graph Transformers for Crystal Material Property Prediction. In *International Conference on Learning Representations*.

Yan, K.; Liu, Y.; Lin, Y.; and Ji, S. 2022. Periodic Graph Transformers for Crystal Material Property Prediction. In *The 36th Annual Conference on Neural Information Processing Systems*.

Zheng, H.; Liu, R.; Lai, F.; and Prakash, A. 2023. Coverage-centric Coreset Selection for High Pruning Rates. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*.

Zhou, J.; Shen, L.; Costa, M. D.; Persson, K. A.; Ong, S. P.; Huck, P.; Lu, Y.; Ma, X.; Chen, Y.; Tang, H.; et al. 2019. 2DMatPedia, an open computational database of two-dimensional materials from top-down and bottom-up approaches. *Scientific data*, 6(1): 86.