PCCP

PAPER

View Article Online
View Journal | View Issue

Cite this: *Phys. Chem. Chem. Phys.*, 2026, **28**, 10119

Comparative assessment of composition- and structure-based surrogate models across 2D materials databases

Inhyo Lee,[a] Hyeokjae Chae,[a] Jongwon Park,[a] Jihye Shin,[a] Hugon Lee*[a] and Seunghwa Ryu *[ab]

Machine learning (ML) surrogate models are increasingly employed to accelerate materials discovery, yet their transferability across heterogeneous databases remains unclear. In this study, we benchmark the accuracy and transferability of composition-based and structure-based surrogate models using three widely adopted two-dimensional (2D) materials databases: Computational 2D Materials Database (C2DB), 2D Materials Encyclopedia (2DMatPedia), and Joint Automated Repository for Various Integrated Simulations (JARVIS-2D). We evaluate predictive performance for energy per atom and bandgap under database-to-database transfer, probe the effects of dataset size and coverage through down-sampling, and analyze error correlations across surrogate models used in this study. Energy-related properties were predicted robustly, whereas bandgap prediction proved to be substantially more difficult due to data imbalance and inconsistencies in DFT parameters across databases. Composition-based models generally exhibited more stable cross-database performance than structure-based models, underscoring that incorporating structural features does not necessarily lead to better generalization. Through down-sampling and error correlation analyses, we demonstrate that cross-database performance is primarily determined by the coverage of the training dataset and that distinct error patterns emerge from differences in the models' feature representation and architecture. Together, this study provides a systematic characterization of surrogate model robustness across 2D materials databases, offering insights into the factors that determine their transferability in materials discovery.

Received 11th December 2025,
Accepted 19th March 2026

DOI: 10.1039/d5cp04814a

rsc.li/pccp

1. Introduction

The advent of machine learning (ML) surrogate models in materials science has reshaped research paradigms by enabling rapid predictions of material properties at a fraction of the cost of trial-and-error experiments or first-principles simulations such as density functional theory (DFT). While first-principles approaches remain indispensable for accuracy, their high computational expense and limited scalability hinder large-scale exploratory searches. Surrogate models mitigate these limitations by offering fast and efficient screening, thereby guiding experimental validation and accelerating the discovery of promising candidates.[1–6] For example, ML has been used to identify vibrationally stable materials[7] and to discover novel cathode candidates with high ion mobility,[8] alongside other applications ranging from the prediction of mechanical and electronic properties to the assessment of thermodynamic stability and synthesizability, underscoring the versatility of ML-driven approaches in materials science.[9–14]

Despite these advances, a central challenge is generalization under distribution shifts. Surrogate models often achieve high accuracy within the training distribution (*i.e.*, in-distribution (ID)) but deteriorate sharply when applied to out-of-distribution (OOD) data.[15] This limitation is critical for materials discovery, as new materials of interest are typically within the OOD domain. Miscalibrated OOD predictions can both waste experimental resources through false positives and obscure important candidates through false negatives.

Accordingly, evaluating surrogate-model performance in genuine OOD settings has become an important topic. Community benchmarks—Matbench,[16] Open Catalyst Project,[17] MatSciML,[18] and JARVIS-Leaderboard[19]—provide valuable testbeds for assessing model accuracy across diverse material properties. However, these benchmarks generally rely on random train/test splits within the same database, which tend to overestimate OOD performance. Structural and chemical redundancy in materials databases often

[a] *Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon, 34141, Republic of Korea. E-mail: hoogon99@kaist.ac.kr, ryush@kaist.ac.kr*
[b] *KAIST InnoCORE PRISM-AI Center, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141, Republic of Korea*

leads to substantial overlap between the training and test sets, making the resulting evaluation insufficiently sensitive to true distribution shifts.[20,21]

To counter this issue, many studies introduce artificial OOD splits. Common strategies include element-exclusion schemes (leave-X-out), sparsity-based splits that hold out low-density regions of the data space (sparse-X or Y-single), and property-range-based splits.[22–25] While these procedures provide controlled stress tests, they still do not fully capture realistic deployment scenarios.[26–28]

Building on these observations, the limitations of existing OOD evaluation approaches can be addressed in two complementary directions. The first is to establish more realistic OOD scenarios. In this regard, evaluating surrogate models across heterogeneous materials databases is practically meaningful for two reasons. First, such cross-database settings directly mirror situations encountered in practical materials discovery, where a surrogate model trained on one database may need to be applied to another for the same property, enabling rapid screening and ranking with only minimal calibration.[29–31] Second, given their inherent differences in compositional distributions and methodological protocols, these databases provide a realistic setting for evaluating the out-of-distribution performance of surrogate models through cross-database testing.[28]

The second direction is to expand OOD evaluation to emerging material families. Most existing benchmarks primarily focus on bulk materials, which limits their relevance for emerging material families. A representative example of such an emerging class is 2D materials, which possess atomic-scale thickness and highly tunable properties and have rapidly become a particularly important focus in materials research.[32–34] Their reduced dimensionality gives rise to unique electronic, optical, and mechanical behaviors—such as high electrical conductivity in 2D–Cu_2Si[35] and exceptional Li storage capacity in CrB_4 and MoB_4 monolayers[36]—making them attractive for applications in energy storage, sensing, and semiconductors.[37–39] The rapid growth of 2D materials research has been accompanied by the increasing use of ML methods for stability, bandgap, and property prediction.[40–42] Consequently, a growing number of 2D materials databases have been compiled to train these ML models. However, the heterogeneity of the existing 2D materials databases raises fundamental questions about how reliably surrogate models transfer across them.

For these reasons, we systematically evaluate surrogate models in a database-to-database transfer setting, where models trained on one database are directly applied to another without fine-tuning. To ensure a meaningful assessment of OOD generalization, we select three 2D materials databases with distinct construction protocols—including their DFT settings, material-generation methods, and compositional coverage (see details in the Methods section). Although many other 2D materials databases exist, such as Alexandria,[43] MC2D[44] and others,[45,46] selected databases offer heterogeneous workflows and coverage to serve as a practical testbed for cross-database transferability. Specifically, we consider the Computational 2D Materials Database (C2DB),[47] 2D Materials Encyclopedia (2DMatPedia),[48] and Joint Automated Repository for Various Integrated Simulations (JARVIS-2D).[49] By comparing representative composition-based and structure-based surrogate models in OOD settings, we address the following questions:

1. How do the dataset size and coverage influence the extrapolation performance of a surrogate model?

2. How do composition- and structure-based approaches differ in their robustness and generalization?

To this end, we evaluate surrogate models on energy per atom and bandgap. Beyond reporting accuracy and transferability, down-sampling and error correlation analyses were employed to examine how training the data size, distributional coverage, and model architecture influence extrapolation performance.

2. Methods

2.1. Overview of the study design

This study was designed to benchmark the database-to-database transferability of surrogate models for 2D materials. As illustrated in Fig. 1, we considered three representative databases—C2DB,[47] 2DMatPedia,[48] and JARVIS-2D[49]—with two target properties, energy per atom and bandgap. All material databases considered were constructed using DFT calculations. For each experiment, surrogate models were trained on one database and then directly applied to predict the properties of materials in another without fine-tuning, thereby reflecting practical scenarios in which surrogate models are applied to the OOD data. Two widely adopted classes of surrogate models—composition based and structure based—were utilized in the study.

Predictive performance was evaluated using two complementary metrics: mean absolute error (MAE) and Spearman correlation. The MAE quantifies absolute predictive accuracy, while the Spearman correlation coefficient measures the degree to which predicted and true values preserve a monotonic ranking. In addition, we performed down-sampling experiments to probe dataset size and coverage effects and error correlation analyses to examine whether different surrogate classes capture complementary aspects of the data. Together, these analyses provide a systematic framework for assessing model robustness under cross-database transfer.

2.2. Databases and target properties

C2DB is a representative high-throughput 2D materials database containing more than 16 000 materials and has expanded through multiple releases (C2DB-2018,[50] C2DB-2021,[51] and C2DB-2022[47]). Earlier versions combined experimentally synthesized materials with hypothetical structures generated *via* lattice decoration, while the most recent release further incorporates crystal diffusion variational autoencoder (CDVAE)-generated materials.[52] Each entry includes a wide set of properties, from stability-related properties and elasticity to magnetism and band structure.

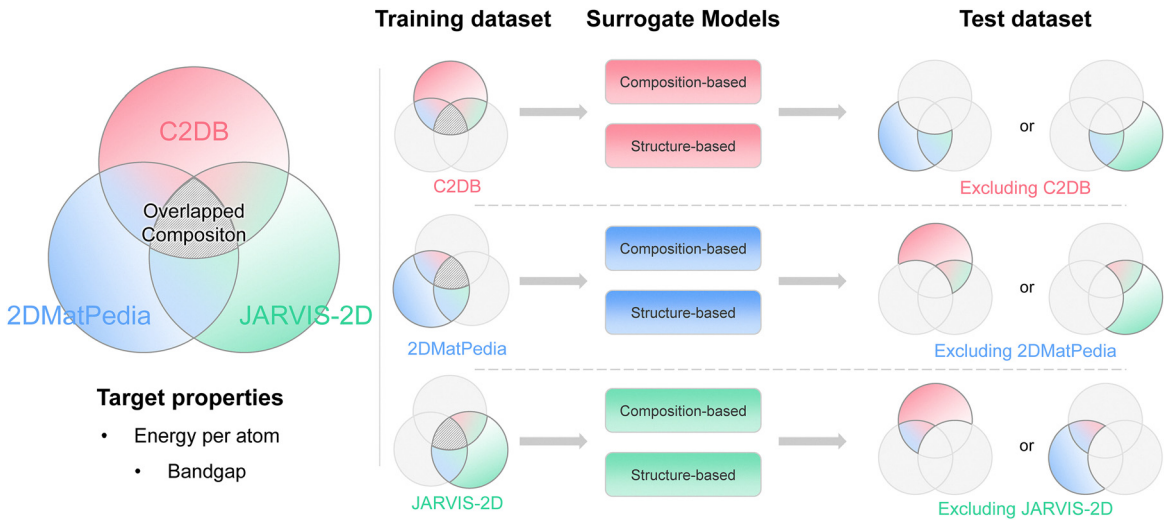

Fig. 1 Database-to-database transfer protocol used in this study. Surrogate models (composition- and structure-based) were trained on one database (C2DB, 2DMatPedia, or JARVIS-2D) and directly tested on another, with overlapping compositions removed from the test dataset to avoid data leakage. Energy per atom and bandgap were considered as target properties.

2DMatPedia comprises more than 6000 materials derived from both exfoliation of layered bulk materials (top-down) and elemental substitution (bottom-up).[48] Reported properties include decomposition energy, exfoliation energy and bandgap, all calculated using DFT.

JARVIS-2D contains about 1100 monolayers generated from a bulk compound using standardized DFT workflows.[49] Alongside formation and exfoliation energies, it reports electronic descriptors such as bandgap, work function, and electron affinity, as well as application-oriented properties such as dielectric properties, Seebeck coefficient, and theoretical solar cell efficiency.

In this study, we consider two targets: energy per atom and bandgap. Energy per atom refers to the ground-state total energy normalized by the number of atoms, serving as the base quantity for deriving thermodynamic stability metrics such as formation energy and energy above the convex hull. The bandgap, defined as the energy difference between the valence and conduction bands, governs a material's conductivity and is therefore a key descriptor for applications in electronics, optoelectronics, and energy devices.

The three databases differ in their exchange-correlation functionals, plane-wave cutoffs, *k*-point sampling densities and others (see details in Section S1 and Table S1, SI), which are known to introduce systematic deviations even for identical materials.[53,54] Such deviations may confound the interpretation of surrogate-model performance under cross-database transfer. To clarify this effect, we compared the target properties for identical materials that appear in multiple databases. While discrepancies arising from methodological differences do shift absolute values, the deviations are generally small, and the overall trends remain consistent (Fig. S1, Tables S2 and S3, SI). For this reason, both MAE and Spearman correlation were employed to evaluate predictive performance.

2.3. Data preprocessing

To ensure comparability across datasets, the following preprocessing rules were applied prior to constructing training and test sets:

1. Entries corresponding to elemental compounds were excluded.

2. For compositions with multiple polymorphs, only the lowest-energy structure was retained.

3. If the same composition appeared in both the training and test sets, the duplicates were removed from the test set.

Although structure-based surrogate models (which will be detailed in Section 2.4) are capable of distinguishing polymorphs with identical compositions, the same preprocessing rules were applied to both composition- and structure-based models to ensure fair comparison. After preprocessing, the dataset sizes were 14 131 (C2DB), 5328 (2DMatPedia), and 893 (JARVIS-2D) for energy per atom, and 6579 (C2DB), 5328 (2DMatPedia), and 893 (JARVIS-2D) for bandgap (Table S4, SI). The distribution of the target properties in the preprocessed datasets is shown in Fig. S2, SI.

2.4. Surrogate models

We considered two classes of surrogate models that are widely used in materials discovery. Composition-based models rely solely on stoichiometric information (*i.e.*, composition), allowing efficient predictions without structural input. Within this class, we examined (i) feature-based models—Random Forest (RF),[55] Gradient Boosted Regression (GBR),[56] and Extra Trees (ET) regressor[57]—which were trained using Magpie descriptors,[58] a fixed set of composition-derived statistical features encoding elemental properties and widely used in materials informatics, and (ii) composition-based neural networks—ElemNet,[59] Representation Learning from Stoichiometry (Roost),[60] and Compositionally Restricted Attention-Based Network (CrabNet)[61]—which

learn embeddings from stoichiometric inputs and iteratively update them through backpropagation.

Structure-based models incorporate structural information such as atomic connectivity and bonding environments,[62–64] typically *via* graph neural networks (GNNs) as the standard approach.[65,66] While often leading to better accuracy, these models typically demand larger training datasets. We employed five widely used architectures: Graph Convolutional Network (GCN),[67] Materials Graph Network (MEGNet),[68] SchNet,[69] Crystal Graph Convolutional Neural Network (CGCNN),[70] and DeeperGATGNN.[71] A brief explanation of each model is provided in Section S3, SI, and hyperparameters are listed in Table S5, SI.

2.5. Experimental settings

We systematically evaluated all possible pairwise transfer cases among the three datasets. For example, models trained on C2DB were tested on both 2DMatPedia and JARVIS-2D, and the process was repeated in the reverse direction (see Fig. 1). Training and test sets were constructed at the composition level to prevent data leakage: overlapping formulas across databases were excluded from the test sets to avoid duplication, even when structural prototypes differed. For structure-based models, full crystal information was used when available, but the same duplicate-removal rules were applied for consistency with composition-based models.

Within-database experiments (training and testing on the same source) served as in-distribution (ID) references. ID evaluation of representative models from each class—feature-based models, composition-based neural networks, and structure-based neural networks—was benchmarked, as summarized in Table 1 for Spearman correlation, showing generally similar performance across the models. Corresponding MAE results are provided in Table S6, SI.

In addition to the database-to-database transfer experiments, we conducted two complementary analyses. First, down-sampling experiments reduced large training sets to the size of smaller databases, enabling the separate assessment of dataset size and distributional coverage. Second, error-correlation analyses quantify similarities and differences in prediction-error patterns across model classes, providing insight into whether different surrogate models capture the complementary aspects of the data. Together, these analyses extend beyond conventional benchmarking and provide deeper understanding of the factors governing model generalization across heterogeneous databases.

Table 1 Spearman correlation results for ID evaluation of three representative surrogate models: RF (tree based), CrabNet (composition based), and DeeperGATGNN (structure based). Each dataset was randomly split into 80% training/validation and 20% independent test sets

		Spearman correlation		
Database	Property	RF	CrabNet	DeeperGATGNN
C2DB	Energy per atom (eV atom^{-1})	0.99	1.00	1.00
	Bandgap (eV)	0.81	0.84	0.82
2DMatPedia	Energy per atom (eV atom^{-1})	0.97	0.98	0.99
	Bandgap (eV)	0.76	0.75	0.73
JARVIS-2D	Energy per atom (eV atom^{-1})	0.95	0.97	0.96
	Bandgap (eV)	0.77	0.74	0.74

3. Results and discussion

In this section, we first compare the distributional characteristics of the databases (Section 3.1), followed by an evaluation of cross-database predictive performance of surrogate models (Section 3.2). We then analyze the effect of dataset size *versus* coverage through down-sampling (Section 3.3) and compare the predictive performance of representative composition- and structure-based models while examining error correlations across all surrogate models considered in this study (Section 3.4).

3.1. Dataset distribution

We begin by analyzing the compositional distributions of the three preprocessed databases to provide context for subsequent transfer experiments. Fig. 2 shows two-dimensional projections of the training data obtained using Uniform Manifold Approximation and Projection (UMAP).[72] Embedded features were represented *via* Magpie descriptors.

For energy per atom, C2DB covers the broadest compositional space (Fig. 2(a)), consistent with its larger dataset size (14 131 materials) and inclusion of multi-component compounds generated by the CDVAE model.[52] Many of these compositions occupy regions not represented in 2DMatPedia (5328 materials; Fig. 2(b)) or JARVIS-2D (893 materials; Fig. 2(c)), both of which are dominated by materials obtained through bulk-derived or bottom-up approaches. This compositional diversity was quantitatively confirmed by calculating the Jensen–Shannon (JS) divergence values on the UMAP projected space. The JS divergence quantifies the similarity between two distributions, with values toward zero indicating greater similarity. Consistent with visual observations, the calculated JS divergences show that 2DMatPedia and JARVIS-2D are most similar (0.05), while C2DB shows a slightly higher divergence from 2DMatPedia (0.11) and JARVIS-2D (0.09).

For bandgap, the coverage of C2DB narrows because bandgaps were calculated only for thermodynamically and dynamically stable materials in the database construction workflow. Nevertheless, C2DB maintains the widest range (Fig. 2(d)), whereas JARVIS-2D exhibits the sparsest (Fig. 2(f)). The corresponding JS divergence values are slightly smaller than those for energy per atom, indicating better similarity across databases: 0.10 for C2DB–2DMatPedia, 0.06 for C2DB–JARVIS-2D, and 0.05 for 2DMatPedia–JARVIS-2D.

Distribution analyses based on UMAP-projected Magpie structural features showed similar trends, with C2DB exhibiting greater structural diversity and JARVIS-2D exhibiting the most restricted coverage (Fig. S3, SI). These differences provide the

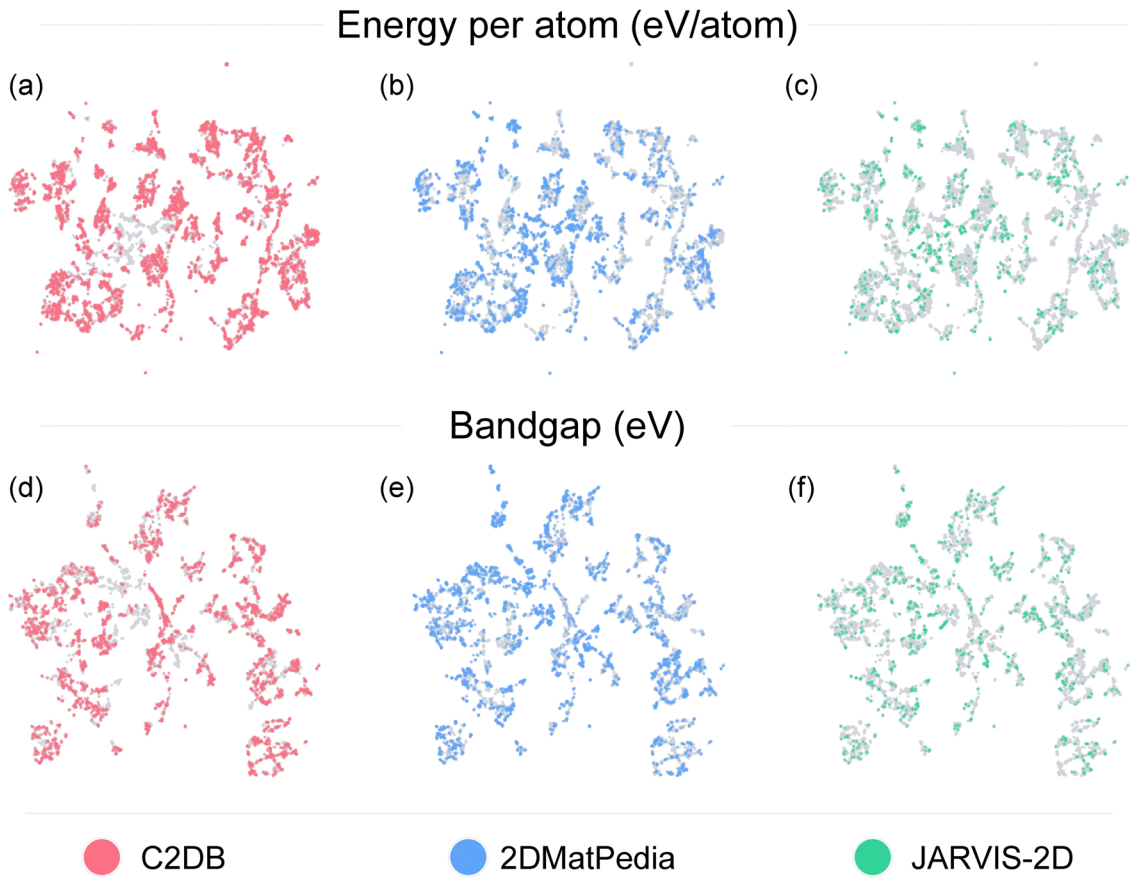

Fig. 2 Distribution of compositions in three 2D materials databases for two target properties *via* UMAP projection. Panels (a)–(c) show the distributions of energy per atom (eV atom^{-1}), and panels (d)–(f) show the distributions of bandgap (eV). Gray points denote the background distribution of the other databases.

basis for evaluating cross-database transfer in the following sections.

3.2. Database-to-database transferability

This section presents predictive performance in database-to-database transfer scenarios. Across all surrogate models, energy per atom showed consistently high transferability, whereas bandgap predictions showed relatively lower accuracy. Fig. 3 summarizes the cross-database performance of composition- and structure-based models.

For energy per atom, Spearman correlations typically range from 0.84 to 0.98, indicating generally consistent results among models and high transferability. The best performance was observed for 2DMatPedia–JARVIS-2D (with averages of 0.95 across composition-based models and 0.96 across structure-based models, Fig. 3(c)). By contrast, models trained on C2DB achieved lower transferability ($\approx$0.85–0.87, Fig. 3(b and c)), despite the larger dataset size. This outcome reflects a distributional mismatch: C2DB contains numerous CDVAE-generated multi-component compositions, which expand the chemical space but limit overlap with 2DMatPedia and JARVIS-2D.

For bandgap prediction, average Spearman correlations ranged from 0.47 to 0.66, indicating substantially lower transferability than for energy per atom. This reduction stems from the well-known, material-dependent errors in DFT bandgap calculations. Self-interaction and the incomplete treatment of electron correlation lead to systematic underestimation, and the magnitude of this error varies across materials due to differences in orbital localization and local bonding environments.[73,74] Additional inconsistencies arise from the empirical choices of Hubbard U parameters used in different databases, further introducing non-uniform biases that surrogate models cannot easily learn.[75] Moreover, the bandgap distribution is highly imbalanced, with many materials clustered near 0 eV and relatively few wide-gap cases (Fig. S2, SI), further complicating training.

Nevertheless, bandgaps from lower-cost DFT approximations (*e.g.*, GGA-PBE[76]) often show strong linear correlations with values obtained from higher-level methods such as HSE06[77] or GW,[78] particularly within specific chemical families. Thus, predicting bandgaps from lower-cost DFT approximations remains valuable, offering an efficient pathway toward more accurate bandgap estimates.[79,80]

Within this overall trend, the highest transferability was observed for 2DMatPedia–C2DB (0.63 for composition-based and 0.60 for structure-based models, Fig. 3(d)), while the lowest was found for JARVIS-2D–2DMatPedia (0.55 and 0.48, respectively, Fig. 3(e)). Unlike energy per atom, C2DB-trained models showed relatively better transferability for bandgap, which can

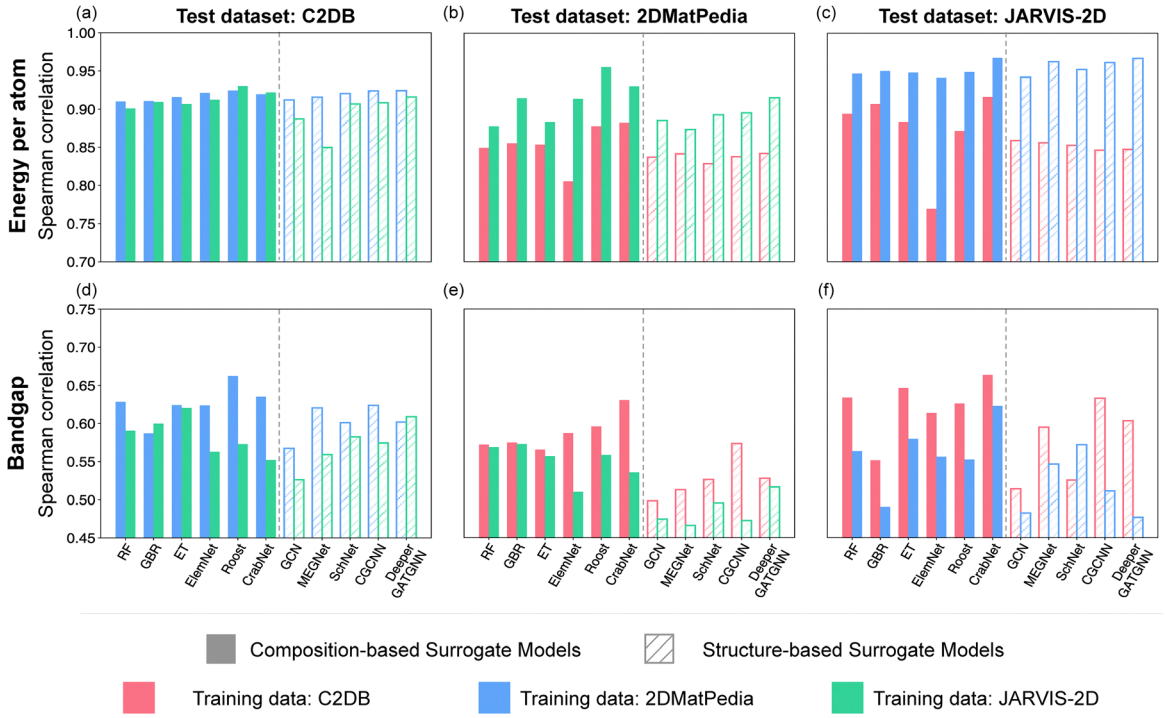

Fig. 3 Database-to-database performance of surrogate models for (a)–(c) energy per atom and (d)–(f) bandgap.

be explained by greater local similarity between C2DB training samples and test samples from other databases (Table S7, SI). The roles of global and local similarity in shaping predictive performance are further discussed in Section 3.3.

At the model level, feature-based models served as stable baselines, with GBR achieving correlations up to 0.95 (Fig. 3(c)). Among neural composition-based models, CrabNet consistently achieved the highest performance, while ElemNet showed the weakest results (*e.g.*, 0.77 for C2DB–JARVIS-2D, Fig. 3(c)). For energy per atom, structure-based models showed broadly similar performance, with DeeperGATGNN and CGCNN typically achieving the highest correlations and GCN the lowest. Nevertheless, composition-based models outperformed structure-based models in terms of overall predictive performance. Higher Spearman correlations were generally associated with lower MAE values. Full MAE and Spearman correlation results are presented in Tables S8 and S9, SI, respectively, for composition-based models, with a visual summary of MAE in Fig. S4, SI. The corresponding results for structure-based models are provided in Tables S10 and S11 and Fig. S5, SI, respectively.

These results demonstrate that (i) predictive performance in database-to-database transfer depends not only on the size of the training data but also on their coverage, and (ii) unlike in ID evaluations, cross-database transfer revealed substantial differences in predictive performance among surrogate models trained on the same dataset.

3.3. Down-sampling: size *versus* coverage

Because database-to-database transfer performance depends strongly on the choice of training data, an important question is whether the observed variability originates from dataset size or from differences in coverage. To assess the effect of training dataset size, we performed down-sampling experiments on C2DB and 2DMatPedia, reducing them to match the sizes of the smaller databases. For each down-sampled case, models were trained on ten random subsets. Fig. 4 shows the results for CrabNet and DeeperGATGNN, which represent the best performing surrogates for composition- and structure-based models, respectively.

For energy per atom, sensitivity to data size depended on the train–test pairing. For CrabNet (Fig. 4(a)), C2DB–2DMatPedia dropped from 0.88 (full) to 0.72 ± 0.03 when down-sampled to the JARVIS-2D size. By contrast, in C2DB–JARVIS-2D, transferability remained nearly unchanged (0.92 full *vs.* 0.90 ± 0.01 down-sampled). When trained on 2DMatPedia, both models transferred relatively well to the other databases (Fig. 4(b)). Similar tendencies were also observed for DeeperGATGNN. For bandgap, down-sampling effects were more pronounced. In C2DB–JARVIS-2D, Spearman correlation decreased from 0.66 to 0.50 ± 0.04 for CrabNet and from 0.60 to 0.50 ± 0.03 for DeeperGATGNN (Fig. 4(c)). In 2DMatPedia–C2DB, performance dropped by approximately 0.08–0.11 for both CrabNet and DeeperGATGNN (Fig. 4(d)). Variance across the ten down-sampled subsets increased as the training size decreased.

To isolate the role of coverage independent of dataset size, we compared equal-sized down-sampled subsets. Fig. 5 illustrates this comparison by showing the compositional feature distributions in the UMAP space, in the case of 2DMatPedia–JARVIS-2D bandgap prediction using CrabNet. It contrasts high- and low-performance training subsets, each down-sampled to

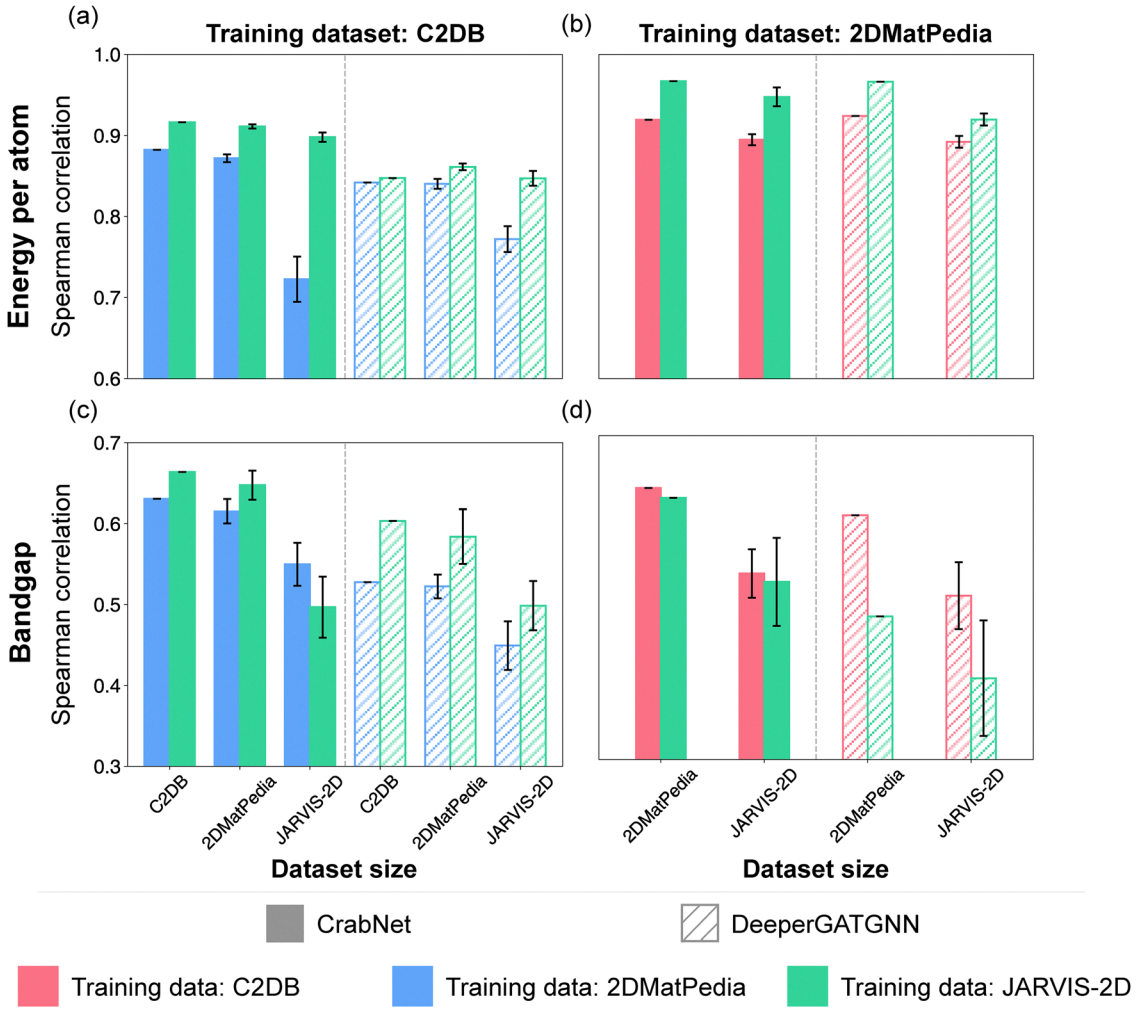

Fig. 4 Database-to-database predictive performance of CrabNet and DeeperGATGNN in down-sampling experiments. Panels (a) and (b) show results for energy per atom, and panels (c) and (d) for bandgap. Models were trained on either C2DB (a) and (c) or 2DMatPedia (b) and (d), using (i) the full dataset, (ii) a subset down-sampled to match the size of 2DMatPedia, and (iii) a subset down-sampled to match the size of JARVIS-2D.

match the size of JARVIS-2D, with the corresponding test data. Nearest-neighbor distances measure the closeness of each test sample to the training data in the UMAP space.

The global distributional similarity in the UMAP space was nearly identical between the two subsets. Both showed comparable JS divergence (0.09 *vs.* 0.10) and average nearest-neighbor distances (0.12 *vs.* 0.13) (Fig. 5(a and b)). In contrast, the local similarity revealed clear differences. Specifically, when examining the ten test samples with the largest nearest-neighbor distances to the training data, the average distance was 0.31 for the high-performance set and 0.38 for the low-performance set (Fig. 5(c)).

These results demonstrate that predictive reliability depends not only on global distributional similarity but also on how well local regions of the test space are covered by the training data. Consequently, reducing the size of the training set diminishes its coverage and leads to degraded transfer performance. Consistently, larger nearest-neighbor distances in the latent space are associated with lower predictive accuracy.[81]

Therefore, variations in coverage across databases naturally give rise to differences in predictive performance. A larger training dataset does not necessarily translate into stronger transferability, since size alone does not ensure distributional similarity with the target data. For instance, although C2DB contains far more entries for energy per atom, its transfer performance to JARVIS-2D or 2DMatPedia is not consistently superior to that of models trained on the smaller databases.

3.4. Comparative analysis and error correlation

Using the full datasets in the database-to-database transfer setting, we observed substantial variability in predictive performance across models (Fig. 3). To analyze this variability, we compare CrabNet and DeeperGATGNN—representatives of the best-performing composition- and structure-based models, respectively—and further examine their error patterns. The transferability results are shown in Fig. 6(a and b), while Fig. 6(c and d) present error-correlation matrices across all surrogate models.

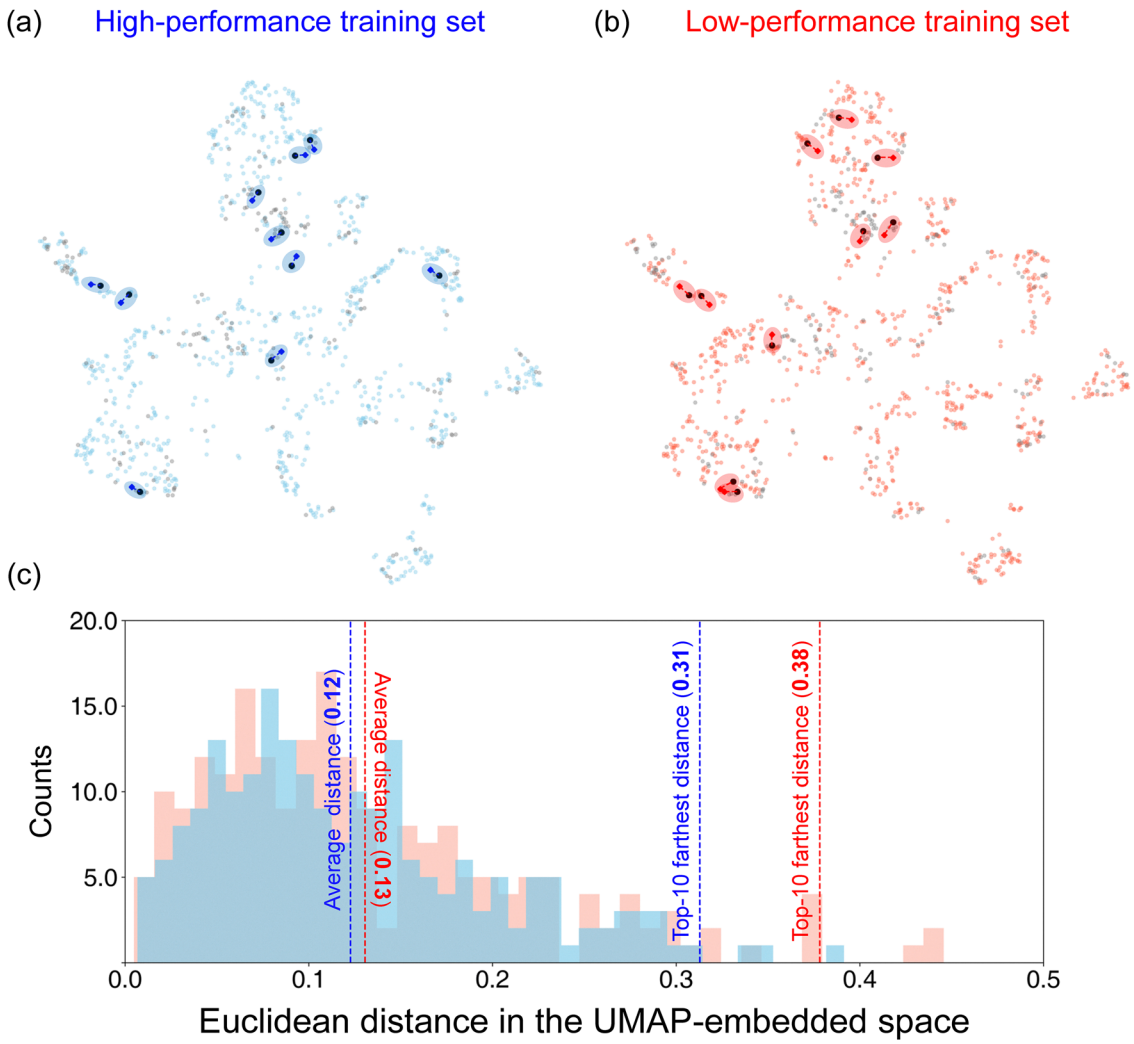

Fig. 5 UMAP projections of the compositional feature space and nearest-neighbor distance analysis for bandgap prediction in the 2DMatPedia–JARVIS-2D transfer, where 2DMatPedia was down-sampled to the size of JARVIS-2D. (a) High-performance training set (blue) and test set (black), with the darkest points indicating the ten samples with the largest nearest-neighbor distances. (b) Low-performance training set (red) and test set (black), highlighting the corresponding distant samples. (c) Comparison of nearest-neighbor distance distributions between training and test sets.

For energy per atom, both models achieved high Spearman correlation (0.84–0.98, Fig. 6(a)). CrabNet slightly outperformed DeeperGATGNN when trained on C2DB (0.92 *vs.* 0.85 for C2DB–JARVIS-2D), whereas the performance was comparable when trained on 2DMatPedia or JARVIS-2D.

For bandgap, the differences were more pronounced (Fig. 6(b)). CrabNet consistently exhibited stronger transferability, outperforming DeeperGATGNN by 0.06–0.10 in C2DB–2DMatPedia and C2DB–JARVIS-2D. The only case favoring DeeperGATGNN was when trained on JARVIS-2D, where it reached 0.61 compared to CrabNet's 0.55.

Overall, these results suggest that simpler composition-based models can be more robust in OOD scenarios, whereas both model classes show comparable predictive performance in ID settings (Table 1), with structure-based models sometimes showing an advantage for certain properties. This trend arises because structure-based models, while benefiting from richer structural information and higher expressiveness in ID tasks, also introduce higher feature dimensionality and greater sensitivity to representation bias, which can exacerbate overfitting under distributional shift. This reflects a broader trade-off, where improving model expressiveness typically enhances ID accuracy but may degrade OOD performance.[22,82] Consequently, in scenarios with distributional shift or limited access to structural information, composition-based models represent a more robust choice.

Error correlation analysis using Pearson correlation matrices (Fig. 6(c and d)) further clarified these systematic differences across model classes. A higher error correlation indicates that two models tend to make similar predictions for the same materials. For energy per atom, feature-based models clustered together with correlations of 0.79–0.85. Composition-based neural networks formed a separate cluster (0.70–0.76), with structure-based models ranging from 0.70 to 0.86. By contrast, cross-family correlations were substantially lower (*e.g.*, RF–SchNet: 0.43), highlighting that different model classes capture complementary aspects of the data.

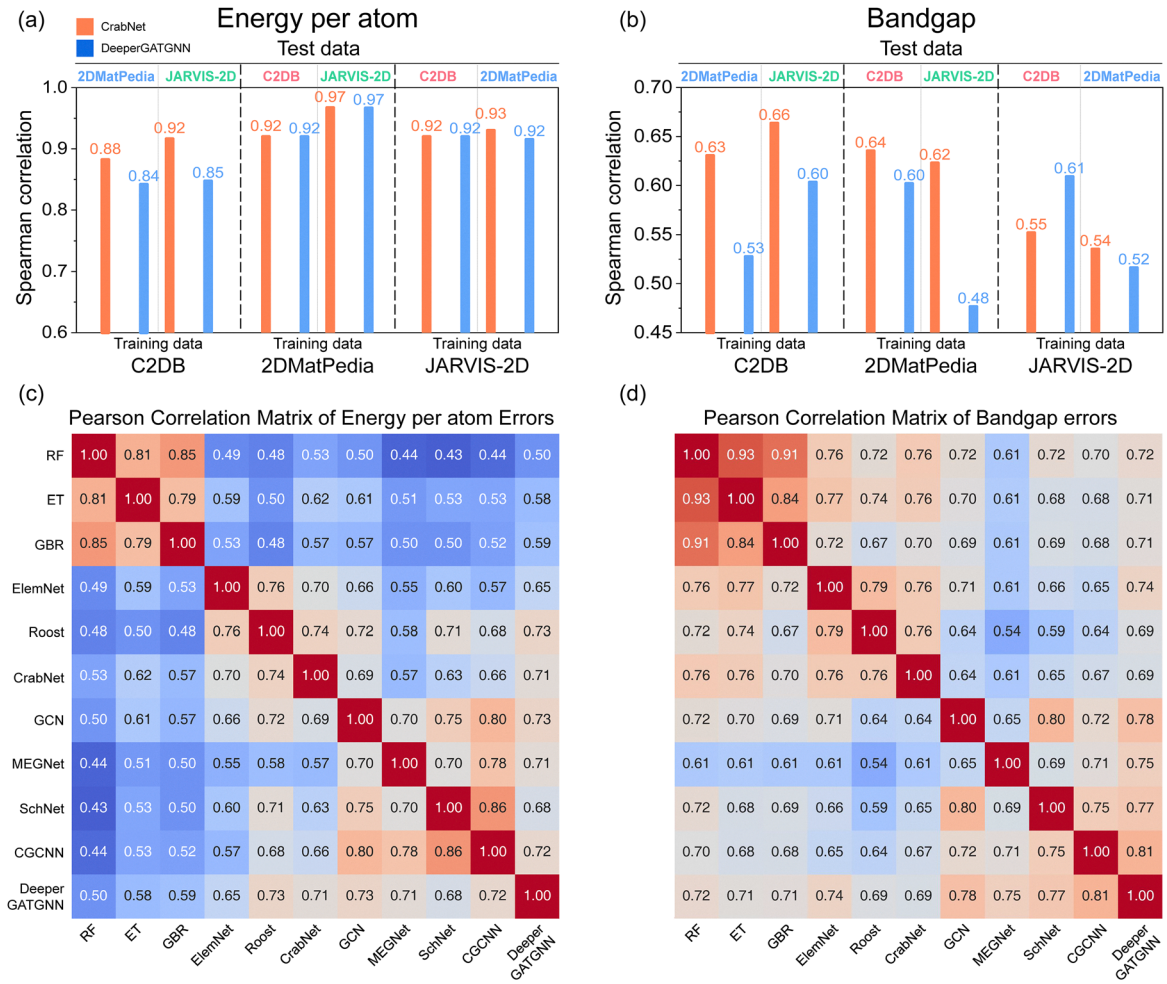

Fig. 6 Comparative analysis of a composition-based model (CrabNet) and a structure-based model (DeeperGATGNN). (a) and (b) Database-to-database Spearman correlation for energy per atom and bandgap predictions, respectively. CrabNet results are shown in orange color and DeeperGATGNN results in blue color. (c) Pearson correlation matrix of prediction errors across all surrogate models for the 2DMatPedia to JARVIS-2D case, which yielded the highest average Spearman correlation between CrabNet and DeeperGATGNN in energy per atom predictions. (d) Pearson correlation matrix of prediction errors for the C2DB to JARVIS-2D case, corresponding to the highest average Spearman correlation between CrabNet and DeeperGATGNN in bandgap predictions.

For bandgap, error correlations were more dispersed. Feature-based models still showed relatively high correlation (0.84–0.93), composition-based neural networks ranged from 0.76 to 0.79, and structure-based neural networks from 0.72 to 0.80, except for MEGNet. Correlations between composition- and structure-based models typically fell in the range of 0.60 to 0.70. The tendency of models within the same class to form distinct error-correlation clusters was even more pronounced in other database transfer scenarios (Fig. S6 and S7, SI).

Feature-based models, which rely on a fixed set of predefined features, naturally exhibited high error correlations. By contrast, neural network models learn internal feature representations during training, leading to lower correlations. In addition, error correlations tended to be stronger within composition- or structure-based classes than across them, reflecting the distinct input information used by each class of models.

Taken together, our analysis shows that predictive performance of surrogate models in cross-database transfer is shaped by two main factors. First, the overall level of predictive performance is largely determined by the distributional similarity between the training and test datasets. Second, even with the same training data, the pattern of predictive error varies with the input representation and architecture of the models. This explains why models that perform similarly in ID evaluations can diverge substantially in OOD settings: each model class relies on different input information and processing strategies—using fixed descriptors, stoichiometric embeddings, or graph-based structural representations—resulting in distinct strengths and limitations in OOD transfer.

4. Conclusion

In this study, we systematically benchmarked the transferability of surrogate models across three widely used 2D material databases: C2DB, 2DMatPedia, and JARVIS-2D. We also conducted down-

sampling experiments and error correlation analyses to further probe the factors governing transfer performance. We considered energy per atom and bandgap as representative target properties. While energy per atom exhibited robust transferability, bandgap prediction remained challenging due to methodological inconsistencies and data imbalance.

Through down-sampling experiments, we confirmed that predictive performance depends not only on dataset size but also on its coverage, including both global and local distributional similarity between training and test sets. Error correlation analyses revealed that models within the same class shared similar error patterns. Even for the same material, differences in how each model family represents the input information led to different error behaviors. Additionally, composition-based models often show more stable performance in out-of-distribution settings than structure-based models.

Overall, this study provides a systematic benchmark for surrogate models by introducing database-to-database transfer as a practical OOD evaluation setting in 2D materials discovery. The findings highlight the importance of dataset coverage, property-specific challenges, and model diversity, offering guidance for building more reliable and transferable ML frameworks in materials science. Future work could extend this framework by incorporating additional databases and target properties, as well as by evaluating a broader class of surrogate models, potentially with tailored architectures.

Conflicts of interest

The authors declare no conflicts of interest.

Data availability

Data will be available upon reasonable request to the corresponding author.

Supplementary information (SI) is available. See DOI: https://doi.org/10.1039/d5cp04814a.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grand funded by the Korean government (MSIT) (No. RS-2023-00222166 and RS-2025-16070951) and the Innocore program of the Ministry of Science and ICT (N10250154).

References

1 J. Sun, J. H. Song, J. Kim, S. Kang, E. Park, S. W. Seo and K. Min, *RSC Adv.*, 2024, **14**, 31439–31450.
2 P.-C. Van Nguyen, V.-T. To, N.-V. Nguyen Tran, T.-L. Phan, T. N. Truong, T. Gärtner, D. Merkle and P. F. Stadler, *Digital Discovery*, 2026, **5**, 241–253.
3 M. Zhu, A. Mroz, L. Gui, K. E. Jelfs, A. Bemporad, E. A. del Río Chanona and Y. S. Lee, *Digital Discovery*, 2024, **3**, 2589–2606.
4 J. Cai, X. Chu, K. Xu, H. Li and J. Wei, *Nanoscale Adv.*, 2020, **2**, 3115–3130.
5 I. Lee, J. Lee, M. Kim, J. Park, H. Kim, S. Lee and K. Min, *ACS Appl. Mater. Interfaces*, 2024, **16**, 52162–52178.
6 J. Lee, I. Lee, J. Park, H. Kim, M. Kim, K. Min and S. Lee, *Chem. Mater.*, 2023, **35**, 10457–10475.
7 S. A. Tawfik, M. Rashid, S. Gupta, S. P. Russo, T. R. Walsh and S. Venkatesh, *npj Comput. Mater.*, 2023, **9**, 5.
8 P. Zhong, B. Deng, T. He, Z. Lun and G. Ceder, *Joule*, 2024, **8**, 1837–1854.
9 S. Kang, M. Kim and K. Min, *J. Phys. Chem. C*, 2023, **127**, 19335–19343.
10 B. Mortazavi, X. Zhuang, T. Rabczuk and A. V. Shapeev, *Mater. Horiz.*, 2023, **10**, 1956–1968.
11 S. Kang, H. Moon, S. Shin, M. Mousavi, H. Sung and S. Ryu, *Mater. Des.*, 2025, **252**, 113798.
12 Y. Wu, J. He, J. Liu, H. Xing, Z. Mao and Y. Liu, *Nanotechnology*, 2018, **30**, 035702.
13 H. Huo, C. J. Bartel, T. He, A. Trewartha, A. Dunn, B. Ouyang, A. Jain and G. Ceder, *Chem. Mater.*, 2022, **34**, 7323–7336.
14 H. Moon, D. Park, H. Cho, H. K. Noh, J. H. Lim and S. Ryu, *Comput. Methods Appl. Mech. Eng.*, 2025, **446**, 118258.
15 K. Li, A. N. Rubungo, X. Lei, D. Persaud, K. Choudhary, B. DeCost, A. B. Dieng and J. Hattrick-Simpers, *Commun. Mater.*, 2025, **6**, 9.
16 A. Dunn, Q. Wang, A. Ganose, D. Dopp and A. Jain, *npj Comput. Mater.*, 2020, **6**, 138.
17 L. Chanussot, A. Das, S. Goyal, T. Lavril, M. Shuaibi, M. Riviere, K. Tran, J. Heras-Domingo, C. Ho, W. Hu, A. Palizhati, A. Sriram, B. Wood, J. Yoon, D. Parikh, C. L. Zitnick and Z. Ulissi, *ACS Catal.*, 2021, **11**, 6059–6072.
18 K. L. K. Lee, C. Gonzales, M. Nassar, M. Galkin and S. Miret, *arXiv*, 2023, arXiv:2309.05934, DOI: 10.48550/arXiv.2309.05934.
19 K. Choudhary, D. Wines, K. Li, K. F. Garrity, V. Gupta, A. H. Romero, J. T. Krogel, K. Saritas, A. Fuhr, P. Ganesh, P. R. C. Kent, K. Yan, Y. Lin, S. Ji, B. Blaiszik, P. Reiser, P. Friederich, A. Agrawal, P. Tiwary, E. Beyerle, P. Minch, T. D. Rhone, I. Takeuchi, R. B. Wexler, A. Mannodi-Kanakkithodi, E. Ertekin, A. Mishra, N. Mathew, M. Wood, A. D. Rohskopf, J. Hattrick-Simpers, S. H. Wang, L. E. K. Achenie, H. Xin, M. Williams, A. J. Biacchi and F. Tavazza, *npj Comput. Mater.*, 2024, **10**, 93.
20 K. Li, B. DeCost, K. Choudhary, M. Greenwood and J. Hattrick-Simpers, *npj Comput. Mater.*, 2023, **9**, 55.
21 J. Hu, D. Liu, N. Fu and R. Dong, *Digital Discovery*, 2024, **3**, 300–312.
22 K. Li, A. N. Rubungo, X. Lei, D. Persaud, K. Choudhary, B. DeCost, A. B. Dieng and J. Hattrick-Simpers, *Commun. Mater.*, 2025, **6**, 9.
23 S. S. Omee, N. Fu, R. Dong, M. Hu and J. Hu, *npj Comput. Mater.*, 2024, **10**, 144.
24 H. Lee, H. Moon, J. Lee and S. Ryu, *Adv. Intell. Discovery*, 2025, e202500107.
25 E. R. Antoniuk, S. Zaman, T. Ben-Nun, P. Li, J. Diffenderfer, B. Sahin, O. Smolenski, T. Hsu, A. M. Hiszpanski, K. Chiu,

B. Kailkhura and B. Van Essen, *arXiv*, 2025, arXiv: 2505.01912, DOI: 10.48550/arXiv.2505.01912.
26 A. F. Zahrt, J. J. Henle and S. E. Denmark, *ACS Comb. Sci.*, 2020, **22**, 586–591.
27 B. Meredig, E. Antono, C. Church, M. Hutchinson, J. Ling, S. Paradiso, B. Blaiszik, I. Foster, B. Gibbons, J. Hattrick-Simpers, A. Mehta and L. Ward, *Mol. Syst. Des. Eng.*, 2018, **3**, 819–825.
28 J. Riebesell, R. E. A. Goodall, P. Benner, Y. Chiang, B. Deng, G. Ceder, M. Asta, A. A. Lee, A. Jain and K. A. Persson, *Nat. Mach. Intell.*, 2025, **7**, 836–847.
29 A. Elrashidy, J. Della-Giustina and J. A. Yan, *J. Phys. Chem. C*, 2024, **128**, 6007–6018.
30 J. Liang, C. Ye, X. Lin, C. Yu, Y. Rouzhahong, C. Liang and H. Li, *J. Phys. Chem. Lett.*, 2025, **16**, 8723–8730.
31 X. Chen, S. Lu, Q. Chen, Q. Zhou and J. Wang, *Nat. Commun.*, 2024, **15**, 5391.
32 H. Kim, C. Kim, Y. Jung, N. Kim, J. Son and G. H. Lee, *Nanotechnology*, 2024, **35**, 262501.
33 D. Wang, S. Ostresh, D. Streater, P. He, J. Nyakuchena, Q. Ma, X. Zhang, J. Neu, G. W. Brudvig and J. Huang, *Angew. Chem. Int. Ed.*, 2023, **62**, e202309505.
34 Y. Wu, J. He, J. Liu, H. Xing, Z. Mao and Y. Liu, *Nanotechnology*, 2018, **30**, 035702.
35 K. M. Yam, N. Guo and C. Zhang, *Nanotechnology*, 2018, **29**, 245704.
36 M. Kashif Masood, J. Wang, J. Song and Y. Liu, *Appl. Surf. Sci.*, 2024, **652**, 159301.
37 X. Huang, C. Liu and P. Zhou, *npj 2D Mater. Appl.*, 2022, **6**, 51.
38 T. Dong, J. Simões and Z. Yang, *Adv. Mater. Interfaces*, 2020, **7**, 1901657.
39 V. Shanmugam, R. A. Mensah, K. Babu, S. Gawusu, A. Chanda, Y. Tu, R. E. Neisiany, M. Försth, G. Sas and O. Das, *Part. Part. Syst. Charact.*, 2022, **39**, 2200031.
40 Y. Q. Huan, Y. Liu, K. E. J. Goh, S. L. Wong and C. S. Lau, *Nanotechnology*, 2021, **32**, 265203.
41 J. Park, M. Kim, H. Kim, J. Lee, I. Lee, H. Park, A. Lee, K. Min and S. Lee, *Phys. Chem. Chem. Phys.*, 2024, **26**, 10769–10783.
42 B. Ryu, L. Wang, H. Pu, M. K. Y. Chan and J. Chen, *Chem. Soc. Rev.*, 2022, **51**, 1899–1925.
43 H.-C. Wang, J. Schmidt, M. A. L. Marques, L. Wirtz and A. H. Romero, *2D Mater.*, 2023, **10**, 035007.
44 D. Campi, N. Mounet, M. Gibertini, G. Pizzi and N. Marzari, *ACS Nano*, 2023, **17**, 11268–11278.
45 Z. Liu, Z. Zhang, X. Liu, M. Yao, X. He, Y. Sun, X. Chen and L. Zhang, *arXiv*, 2025, arXiv:2507.00584, DOI: 10.48550/arXiv.2507.00584.
46 F. Bertoldo, S. Ali, S. Manti and K. S. Thygesen, *npj Comput. Mater.*, 2022, **8**, 56.
47 P. Lyngby and K. S. Thygesen, *npj Comput. Mater.*, 2022, **8**, 232.
48 J. Zhou, L. Shen, M. D. Costa, K. A. Persson, S. P. Ong, P. Huck, Y. Lu, X. Ma, Y. Chen, H. Tang and Y. P. Feng, *Sci. Data*, 2019, **6**, 86.
49 K. Choudhary, K. F. Garrity, A. C. E. Reid, B. DeCost, A. J. Biacchi, A. R. Hight Walker, Z. Trautt, J. Hattrick-Simpers, A. G. Kusne, A. Centrone, A. Davydov, J. Jiang, R. Pachter, G. Cheon, E. Reed, A. Agrawal, X. Qian, V. Sharma, H. Zhuang, S. V. Kalinin, B. G. Sumpter, G. Pilania, P. Acar, S. Mandal, K. Haule, D. Vanderbilt, K. Rabe and F. Tavazza, *npj Comput. Mater.*, 2020, **6**, 173.
50 S. Haastrup, M. Strange, M. Pandey, T. Deilmann, P. S. Schmidt, N. F. Hinsche, M. N. Gjerding, D. Torelli, P. M. Larsen, A. C. Riis-Jensen, J. Gath, K. W. Jacobsen, J. J. Mortensen, T. Olsen and K. S. Thygesen, *2D Mater.*, 2018, **5**, 042002.
51 M. N. Gjerding, A. Taghizadeh, A. Rasmussen, S. Ali, F. Bertoldo, T. Deilmann, N. R. Knøsgaard, M. Kruse, A. H. Larsen, S. Manti, T. G. Pedersen, U. Petralanda, T. Skovhus, M. K. Svendsen, J. J. Mortensen, T. Olsen and K. S. Thygesen, *2D Mater.*, 2021, **8**, 044002.
52 T. Xie, X. Fu, O.-E. Ganea, R. Barzilay and T. Jaakkola, Crystal Diffusion Variational Autoencoder for Periodic Material Generation, ICLR 2022 – 10th International Conference on Learning Representations.
53 V. I. Hegde, C. K. H. Borg, Z. Del Rosario, Y. Kim, M. Hutchinson, E. Antono, J. Ling, P. Saxe, J. E. Saal and B. Meredig, *Phys. Rev. Mater.*, 2023, **7**, 053805.
54 J. Nisar, C. Århammar, E. Jämstorp and R. Ahuja, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2011, **84**, 075120.
55 L. Breiman, *Mach. Learn.*, 2001, **45**, 5–32.
56 J. H. Friedman, *Annu. Stat.*, 2001, **29**, 1189–1231.
57 P. Geurts, D. Ernst and L. Wehenkel, *Mach. Learn.*, 2006, **63**, 3–42.
58 L. Ward, A. Agrawal, A. Choudhary and C. Wolverton, *npj Comput. Mater.*, 2016, **2**, 16028.
59 D. Jha, L. Ward, A. Paul, W.-K. Liao, A. Choudhary, C. Wolverton and A. Agrawal, *Sci. Rep.*, 2018, **8**, 17593.
60 R. E. A. Goodall and A. A. Lee, *Nat. Commun.*, 2020, **11**, 6280.
61 A. Y. T. Wang, S. K. Kauwe, R. J. Murdock and T. D. Sparks, *npj Comput. Mater.*, 2021, **7**, 77.
62 P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer and P. Friederich, *Commun. Mater.*, 2022, **3**, 93.
63 Y. Zuo, C. Chen, X. Li, Z. Deng, Y. Chen, J. Behler, G. Csányi, A. V. Shapeev, A. P. Thompson, M. A. Wood and S. P. Ong, *J. Phys. Chem. A*, 2020, **124**, 731–745.
64 J. Carrete, W. Li, N. Mingo, S. Wang and S. Curtarolo, *Phys. Rev. X*, 2014, **4**, 011019.
65 V. Fung, J. Zhang, E. Juarez and B. G. Sumpter, *npj Comput. Mater.*, 2021, **7**, 84.
66 P. Reiser, M. Neubert, A. Eberhard, L. Torresi, C. Zhou, C. Shao, H. Metni, C. van Hoesel, H. Schopmans, T. Sommer and P. Friederich, *Commun. Mater.*, 2022, **3**, 93.
67 T. N. Kipf and M. Welling, Semi-Supervised Classification with Graph Convolutional Networks, 5th International Conference on Learning Representations, ICLR 2017 – Conference Track Proceedings.
68 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chem. Mater.*, 2019, **31**, 3564–3572.
69 K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko and K.-R. Müller, *Adv. Neural Inf. Process. Syst.*, 2017, 992–1002.

70 T. Xie and J. C. Grossman, *Phys. Rev. Lett.*, 2018, **120**, 145301.
71 S. S. Omee, S. Y. Louis, N. Fu, L. Wei, S. Dey, R. Dong, Q. Li and J. Hu, *Patterns*, 2022, **3**, 100491.
72 L. McInnes, J. Healy and J. Melville, *arXiv*, 2018, arXiv:1802.03426, DOI: 10.48550/arXiv.1802.03426.
73 A. J. Cohen, P. Mori-Sánchez and W. Yang, *Science*, 2008, **321**, 792–794.
74 T. Tsuneda and K. Hirao, *J. Chem. Phys.*, 2014, **140**, 18A513.
75 P. Borlido, T. Aull, A. W. Huran, F. Tran, M. A. L. Marques and S. Botti, *J. Chem. Theory Comput.*, 2019, **15**, 5069–5079.
76 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865–3868.
77 A. V. Krukau, O. A. Vydrov, A. F. Izmaylov and G. E. Scuseria, *J. Chem. Phys.*, 2006, **125**, 224106.
78 M. A. L. Marques, J. Vidal, M. J. T. Oliveira, L. Reining and S. Botti, *Phys. Rev. B: Condens. Matter Mater. Phys.*, 2011, **83**, 035119.
79 P. Borlido, J. Schmidt, A. W. Huran, F. Tran, M. A. L. Marques and S. Botti, *npj Comput. Mater.*, 2020, **6**, 96.
80 Á. Morales-García, R. Valero and F. Illas, *J. Phys. Chem. C*, 2017, **121**, 18862–18866.
81 J. P. Janet, C. Duan, T. Yang, A. Nandy and H. J. Kulik, *Chem. Sci.*, 2019, **10**, 7913–7922.
82 L. M. Ghiringhelli, J. Vybiral, S. V. Levchenko, C. Draxl and M. Scheffler, *Phys. Rev. Lett.*, 2015, **114**, 105503.

