Supporting Information for

Comparative Assessment of Composition- and Structure-Based Surrogate Models Across 2D Materials Databases

Inhyo Lee[1], *Hyeokjae Chae*[1], *Jongwon Park*[1], *Jihye Shin*[1], *Hugon Lee*[1*], and *Seunghwa Ryu*[1,2*]

[1]Department of Mechanical Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291 Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea

[2]KAIST InnoCORE PRISM-AI Center, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, 34141, Republic of Korea

[*] Corresponding author: hoogon99@kaist.ac.kr (H. Lee) and ryush@kaist.ac.kr (S. Ryu)

S1. Computational settings of 2D databases

The three databases employed in this study were generated using different DFT settings. The details are summarized below.

Table S1. DFT calculation settings used in C2DB, 2DMatPedia, and JARVIS-2D

Categories	C2DB	2DMatPedia	JARVIS-2D
Software	GPAW	VASP	VASP
Pseudopotentials	Projector-augmented wave (PAW)	Projector-augmented wave (PAW)	Projector-augmented wave (PAW)
Hubbard *U*	X	O	O
XC Functionals (Structure optimization)	GGA-PBE	vdW-OptB88	vdW-OptB88
XC Functionals (Electronic properties)	GGA-PBE (for data used in this study)	GGA-PBE (for data used in this study)	vdW-OptB88 (for data used in this study)
Plane-wave cutoff	800 eV	520 eV	Convergence test from 500 eV in 50 eV steps [1]
k-mesh density	6 $Å^{-1}$ (structure optimization), 12 $Å^{-1}$ (Electronic properties)	1000 (number of atoms in the cell)	Convergence test from line-density 0 $Å^{-1}$ in 5 $Å^{-1}$ steps [1]
Vacuum	15 Å	20 Å	20 Å

Differences in DFT parameters—including exchange–correlation functionals, plane-wave cutoffs, *k*-mesh densities, and Hubbard U—can lead to systematic variations in the computed properties even for identical materials. vdW-optB88, for example, includes explicit dispersion interactions and therefore yields different total energies and cell parameters compared with GGA-PBE.[2,3] Numerical settings such as plane-wave cutoff and *k*-mesh density further influence the accuracy of total energy and forces.

To account for these differences in computational methods, we compared the calculated energy per atom and bandgap values for identical materials across the three databases. **Figure S1** presents the distributions of energy per atom and bandgap for each database. For visual

clarity, we annotated only a randomly selected subset of data points from each database; all labels reference the corresponding 2DMatPedia entries for consistency. **Table S2** summarizes the statistical comparison between databases, providing p-values and Spearman correlation coefficients for the overlapping materials. **Table S3** lists the material IDs together with their corresponding energy per atom and bandgap values as reported in each database.

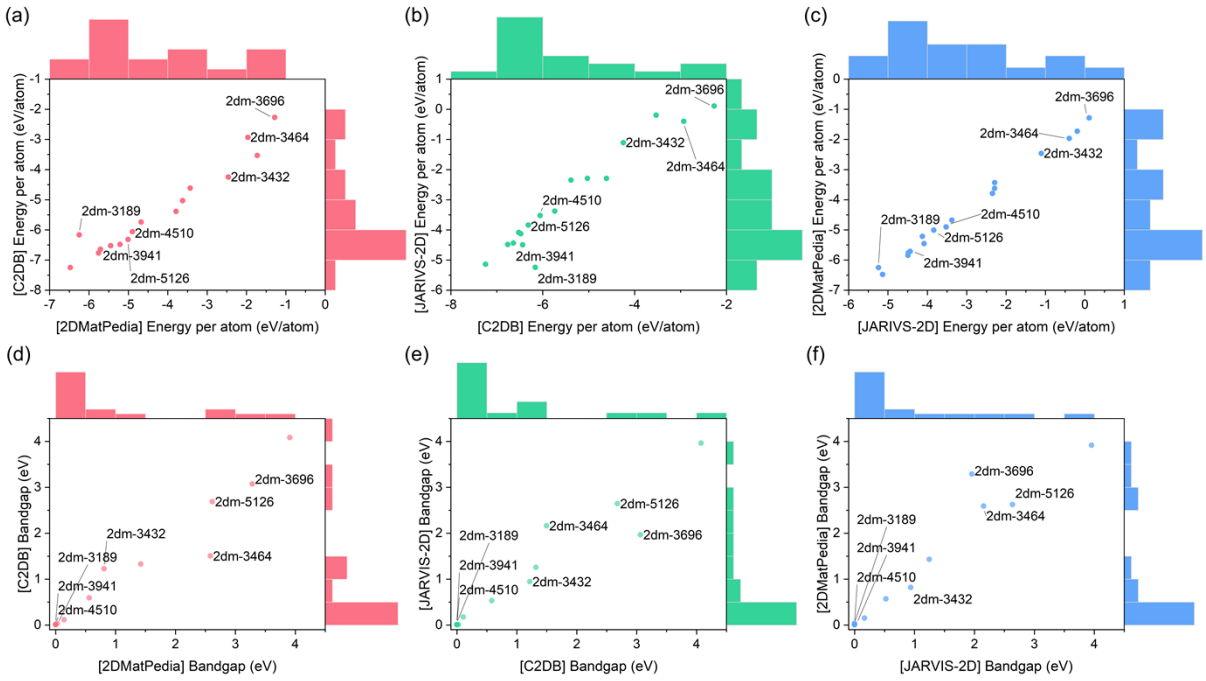

Figure S1. Comparison of calculated property values for identical materials across databases. (a) Energy per atom between C2DB and 2DMatPedia. (b) Energy per atom between C2DB and JARVIS-2D. (c) Energy per atom between 2DMatPedia and JARVIS-2D. (d) Bandgap between C2DB and 2DMatPedia. (e) Bandgap between C2DB and JARVIS-2D. (f) Bandgap between 2DMatPedia and JARVIS-2D. In all panels, only a randomly selected subset of data points is annotated for clarity.

Table S2. Statistical comparison of identical materials across databases. The table reports p-values and Spearman correlation coefficients for energy per atom and bandgap between each pair of databases (C2DB—2DMatPedia, C2DB—JARVIS-2D, and 2DMatPedia—JARVIS-2D)

Property	**Database 1**	**Database 2**	p**-value**	**Spearman correlation**
Energy per atom (eV/atom)	2DMatPedia	C2DB	0.000	0.932
	C2DB	JARVIS-2D	0.000	0.915
	JARVIS-2D	2DMatPedia	0.000	0.996
Bandgap (eV)	2DMatPedia	C2DB	0.000	1.000
	C2DB	JARVIS-2D	0.000	0.966
	JARVIS-2D	2DMatPedia	0.000	0.966

Table S3. List of identical materials included in multiple databases. The table provides material IDs and the corresponding values of energy per atom and bandgap as reported in C2DB, 2DMatPedia, and JARVIS-2D.

Formula	Material ID			Energy per atom (eV/atom)			Bandgap (eV)		
	C2DB	2DMatPedia	JARVIS-2D	C2DB	2DMatPedia	JARVIS-2D	C2DB	2DMatPedia	JARVIS-2D
$CdBr_2$	2dm-3696	1CdBr2-1	JVASP-76511	-2.27	-1.28	0.11	3.07	3.29	1.96
CoS_2	2dm-4324	1CoS2-1	JVASP-28291	-5.74	-4.68	-3.38	0	0	0
$CoTe_2$	2dm-3401	1CoTe2-1	JVASP-77587	-5.03	-3.62	-2.29	0	0	0
FeS_2	2dm-4510	1FeS2-1	JVASP-28292	-6.06	-4.90	-3.52	0	0	0
MnS_2	2dm-4839	1MnS2-1	JVASP-77582	-6.52	-5.45	-4.08	0	0	0
$MoSe_2$	2dm-3409	1MoSe2-1	JVASP-76621	-6.76	-5.76	-4.48	1.32	1.42	1.25
NbS_2	2dm-3019	1NbS2-1	JVASP-60407	-7.25	-6.48	-5.14	0	0	0
$NbSe_2$	2dm-3941	1NbSe2-1	JVASP-76622	-6.64	-5.71	-4.44	0	0	0
$NiCl_2$	2dm-3432	1NiCl2-1	JVASP-76342	-4.25	-2.46	-1.11	1.22	0.81	0.94
NiI_2	2dm-3692	1NiI2-1	JVASP-27920	-3.53	-1.73	-0.19	0.11	0.14	0.17
NiS_2	2dm-5509	1NiS2-1	JVASP-28295	-5.38	-3.79	-2.35	0.58	0.56	0.52
PbI_2	2dm-3464	1PbI2-1	JVASP-76581	-2.93	-1.97	-0.40	1.50	2.59	2.16
RhO_2	2dm-6389	1RhO2-1	JVASP-77581	-6.48	-5.21	-4.13	0	0	0
SnF_4	2dm-3719	1SnF4-1	JVASP-78033	-4.61	-3.43	-2.29	4.08	3.91	3.96
SnO_2	2dm-5126	1SnO2-1	JVASP-28296	-6.32	-5.01	-3.84	2.68	2.62	2.64
VS_2	2dm-3214	1VS2-1	JVASP-75044	-6.44	-5.84	-4.49	0.03	0.03	0
Y_2C	2dm-3189	1CY2-1	JVASP-27865	-6.16	-6.25	-5.24	0	0	0

S2. Database size and property distributions

Table S4. Number of training data samples available in each database after preprocessing for energy per atom and bandgap. C2DB, 2DMatPedia, and JARVIS-2D differ substantially in size, with C2DB providing the largest dataset and JARVIS-2D the smallest.

Database	Training data for energy per atom	Training data for bandgap
C2D8	14,131	6,579
2DMatPedia	5,328	5,328
JARVIS-2D	893	893

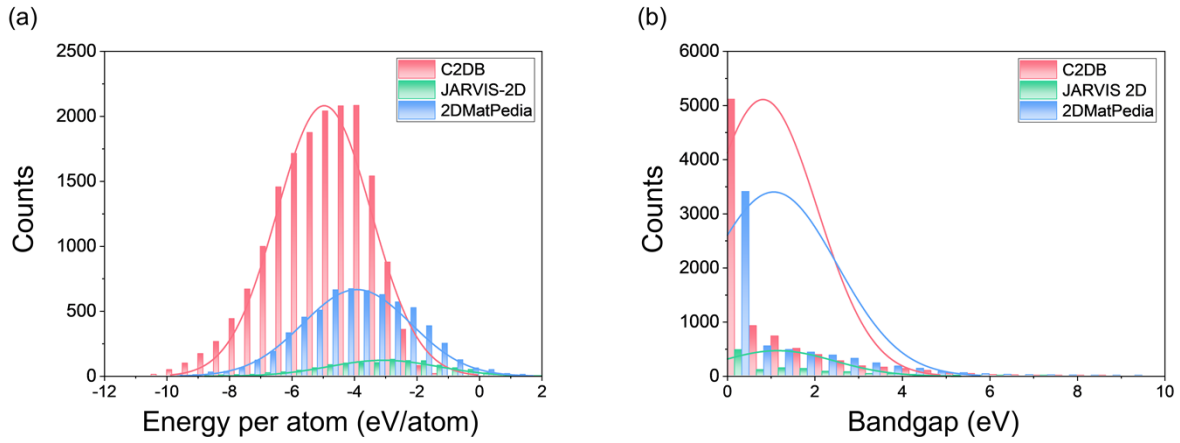

Figure S2. Distribution of target properties in the three databases after preprocessing. (a) Histogram of energy per atom (eV/atom) for C2DB, 2DMatPedia, and JARVIS-2D. (b) Histogram of bandgap (eV) for C2DB, 2DMatPedia, and JARVIS-2D.

S3. Details for data preprocessing and surrogate models

S3.1. Dataset construction rationale

To enable a controlled comparison between composition-based and structure-based surrogate models, the training datasets were constructed such that each composition corresponds to a unique target value. Since composition-only models cannot distinguish between different polymorphs sharing the same chemical formula, retaining multiple structural variants with distinct target values would introduce label ambiguity and render the supervised learning problem ill-posed for compositional models. Therefore, duplicate compositions were removed and a single representative structure per composition was retained. Specifically, the lowest-energy configuration was selected, which is consistent with thermodynamic stability–driven materials discovery workflows.

S3.2. Composition-based models

Random Forest (RF) An ensemble of decision trees built on bootstrapped subsets of data with random feature selection. Predictions are averaged across the trees to reduce variance and mitigate overfitting.[4]

Gradient Boosted Regression (GBR) A boosting method where trees are trained sequentially, each correcting the residual errors of the previous one. This framework is effective at capturing complex, nonlinear relationships between compositional features and target properties.[5]

Extra Trees (ET) An ensemble of decision trees similar to RF, but using randomized features and threshold splits without bootstrapping. This approach increases computational efficiency and reduces variance, though it can introduce higher bias.[6]

ElemNet A deep fully connected neural network trained directly on normalized elemental compositions. It is designed to learn composition–property relationships without relying on predefined, hand-crafted features.[7]

Representation Learning from Stoichiometry (Roost) A graph neural network that represents compositions as element–stoichiometry graphs. It uses message passing and attention pooling to capture element–element interactions and their relative proportions.[8]

Compositionally Restricted Attention-Based network (CrabNet) A transformer-based architecture that encodes compositions as sequences of elemental embeddings with attention

layers. This enables the model to learn higher-order interactions among elements and their stoichiometric ratios without explicit structural information.[9]

S3.3. Structure-based models

Graph Convolutional Network (GCN) A graph neural network where atoms are nodes and edges are typically defined by distance-based neighbor criteria. Node features are updated through the convolutional aggregation of neighbor information.[10]

Materials Graph Network (MEGNet) Extends standard message passing to include atom, bond, and global state variables. This flexible representation allows for the simultaneous learning of both structural and compositional effects.[11]

SchNet A continuous-filter convolutional neural network that embeds interatomic distances through continuous filter functions. It learns quantum chemistry-inspired representations while preserving rotational and translational invariance.[12]

Crystal Graph Convolution Neural Networks (CGCNN) Constructs graphs directly from crystal structures, defining atoms as nodes and distance-based neighbors as edges. Iterative message passing captures local chemical environments and bonding characteristics.[13]

DeeperGATGNN A deep architecture based on Graph Attention Networks. It incorporates attention-based message passing and residual connections to effectively model complex structural dependencies while mitigating oversmoothing in deep GNNs.[14]

S3.4. Hyperparameters

Unless otherwise noted, we used library defaults. Specifically, feature-based models followed the default settings in **PyCaret**, structure-based models followed **MatDeepLearn** defaults, and composition-based neural networks used the default configurations of their reference implementations. To ensure a fair cross-dataset comparison, we standardized only the number of epochs and batch size. All other optimization settings (optimizer, learning rate, weight decay, etc.) were left at their respective defaults (see **Table S5**).

Table S5. Hyperparameters of surrogate models.

Category	Model	Hyperparameters
Feature-based models	RF	n_estimators=100, max_depth=None, max_features=1.0, bootstrap=True, criterion=squared_error, min_samples_split=2, min_samples_leaf=1
	GBR	n_estimators=100, max_depth=3, max_features=None, subsample=1.0, criterion=friedman_mse, loss=squared_error
	ET	n_estimators=100, max_depth=None, max_features=1.0, bootstrap=False, criterion=squared_error, min_samples_split=2, min_samples_leaf=1
Composition-based neural networks	ElemNet	layers=[1024×4, 512×3, 256×3, 128×3, 64×2, 32×1], activation=ReLU, output=1
	Roost	elem_fea_len=64, n_graph=3, elem_heads=3, elem_gate=[256], elem_msg=[256], cry_heads=3, cry_gate=[256], cry_msg=[256], trunk_hidden=[1024,512], out_hidden=[256,128,64], activation=ReLU, dropout=0.0
	CrabNet	d_model=512, L=3, H=4, FFN=2048, dropout=0.1, element_embed=mat2vec→Linear(d_model), fractional_encodings=sinusoidal(linear+log), resolution=5000, output_mlp=[1024,512,256,128], aggregation=masked_mean, gating=sigmoid
Structure-based surrogate models	GCN	dim=[64,64], pre_fc_count=1, gc_count=3, post_fc_count=1, pool=global_mean_pool, pool_order=early, batch_norm=True, activation=ReLU, dropout=0.0
	MEGNet	dim=[64,64,64], pre_fc_count=1, gc_count=3, gc_fc_count=2, post_fc_count=1, pool=global_mean_pool, pool_order=early, batch_norm=True, activation=ReLU, dropout=0.0
	SchNet	dim=[64,64,64], cutoff=8 Å, pre_fc_count=1, gc_count=3, post_fc_count=1, pool=global_mean_pool, pool_order=early, batch_norm=True, activation=ReLU, dropout=0.0
	CGCNN	dim=[64,64], pre_fc_count=1, gc_count=3, post_fc_count=1, pool=global_mean_pool, pool_order=early, batch_norm=True, activation=ReLU, dropout=0.0
	DeeperGATGNN	dim=[64,64], pre_fc_count=1, gc_count=5, post_fc_count=1, pool=global_add_pool, pool_order=early, batch_norm=True, activation=Softplus, dropout=0.0, attention_heads=4, global_attention=True, norm_layer=DiffGroupNorm

S4. In-distribution performance

Table S6. MAE 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.

Database	**Property**	**MAE**		
		RF	**CrabNet**	**DeeperGATGNN**
C2DB	Total energy per atom (eV/atom)	0.15	0.07	0.07
	Bandgap (eV)	0.35	0.27	0.28
2DMatPedia	Total energy per atom (eV/atom)	0.28	0.18	0.16
	Bandgap (eV)	0.56	0.48	0.53
JARVIS-2D	Total energy per atom (eV/atom)	0.39	0.31	0.34
	Bandgap (eV)	0.61	0.59	0.61

S5. Distribution of structural spaces

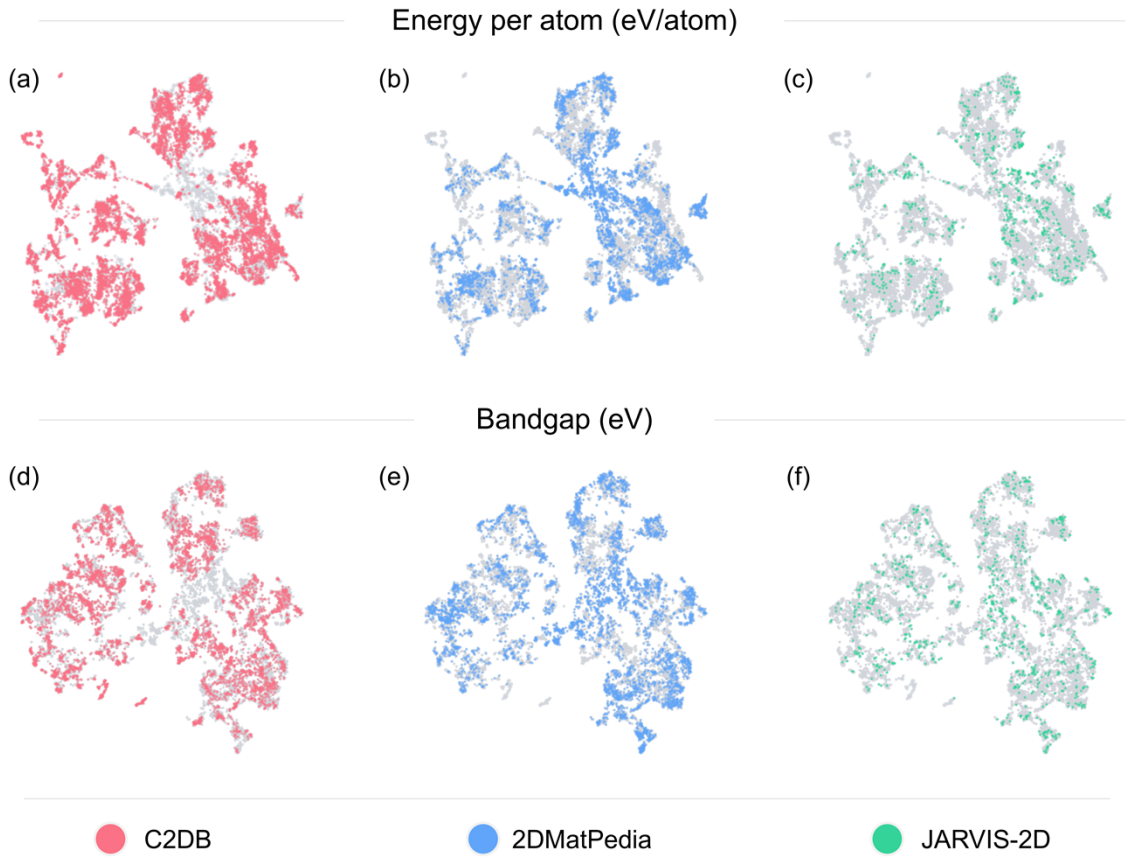

Figure S3. Distribution of structures in three 2D material databases for two target properties. (a)–(c) show the distributions of energy per atom (eV/atom), and panels (d)–(f) show the distributions of bandgap (eV). Each column corresponds to a database: (a) and (d) C2DB (red), (b) and (e) 2DMatPedia (blue), and (c) and (f) JARVIS-2D (green). Gray points denote the background distribution of the other databases, while colored points highlight the corresponding database.

S6. Local similarity of bandgap data between databases

Local similarity between databases was quantified by computing the average distance between each test sample and its k nearest training samples (k = 1, 3, 5, and 10) in the UMAP-projected compositional feature space. Considering multiple neighbors (larger k) reduces sensitivity to individual points and better reflects how well the local region of a test sample is represented in the training data. The reported values correspond to the average over the 10 test samples with the farthest nearest-neighbor distances.

Table S7. Top-k nearest-neighbor distances (k = 1, 3, 5, 10) in the UMAP-projected compositional feature space for bandgap. Reported values are averaged over the 10 test samples with the farthest nearest-neighbor distances, representing local similarity between the three 2D materials databases.

Database 1	Database 2	Top-1	Top-3	Top-5	Top-10
C2DB	2DMatPedia	1.34	1.44	1.55	3.03
C2DB	JARVIS-2D	1.43	1.50	1.98	3.47
JARVIS-2D	2DMatPedia	1.28	2.25	2.43	3.62

S7. Predictive performance of composition-based surrogate models

Table S8. MAE (unitless) of composition-based surrogate models for energy per atom and bandgap under database-to-database transfer settings. Results are reported for RF, GBR, ET, ElemNet, Roost, and CrabNet across all database pairs (C2DB, 2DMatPedia, and JARVIS-2D).

Surrogate model	Property	C2DB to 2DMatPedia	C2DB to JARVIS-2D	2DMatPedia to C2DB	2DMatPedia to JARVIS-2D	JARVIS-2D to C2DB	JARVIS-2D to 2DMatPedia
RF	Energy per atom (eV/atom)	1.15	1.90	0.87	1.23	1.89	1.11
GBR		1.14	1.88	0.88	1.17	1.95	1.09
ET		1.17	1.91	0.87	1.21	1.94	1.09
ElemNet		1.20	1.88	0.88	1.28	1.99	1.09
Roost		1.10	1.81	0.82	1.32	2.01	1.11
CrabNet		1.14	1.99	0.88	1.20	1.99	1.08
RF	Bandgap (eV)	0.82	0.72	0.61	0.72	0.75	0.90
GBR		0.81	0.78	0.66	0.78	0.74	0.89
ET		0.80	0.69	0.61	0.69	0.71	0.89
ElemNet		0.75	0.72	0.62	0.73	0.72	0.87
Roost		0.71	0.66	0.53	0.69	0.68	0.85
CrabNet		0.65	0.67	0.57	0.62	0.70	0.82

Table S9. Spearman correlation of composition-based surrogate models for energy per atom and bandgap under database-to-database transfer settings. Results are reported for RF, GBR, ET, ElemNet, Roost, and CrabNet across all database pairs (C2DB, 2DMatPedia, and JARVIS-2D).

Surrogate model	**Property**	**C2DB to 2DMatPedia**	**C2DB to JARVIS-2D**	**2DMatPedia to C2DB**	**2DMatPedia to JARVIS-2D**	**JARVIS-2D to C2DB**	**JARVIS-2D to 2DMatPedia**
RF	Energy per atom (eV/atom)	0.85	0.89	0.91	0.95	0.90	0.88
GBR		0.86	0.91	0.91	0.95	0.91	0.91
ET		0.85	0.88	0.92	0.95	0.91	0.88
ElemNet		0.81	0.77	0.92	0.94	0.91	0.91
Roost		0.88	0.87	0.92	0.95	0.9.	0.96
CrabNet		0.88	0.92	0.92	0.97	0.92	0.93
RF	Bandgap (eV)	0.57	0.63	0.63	0.56	0.59	0.57
GBR		0.57	0.55	0.59	0.49	0.60	0.57
ET		0.57	0.65	0.62	0.58	0.62	0.56
ElemNet		0.59	0.61	0.62	0.56	0.56	0.51
Roost		0.60	0.63	0.66	0.55	0.57	0.56
CrabNet		0.63	0.66	0.64	0.62	0.55	0.54

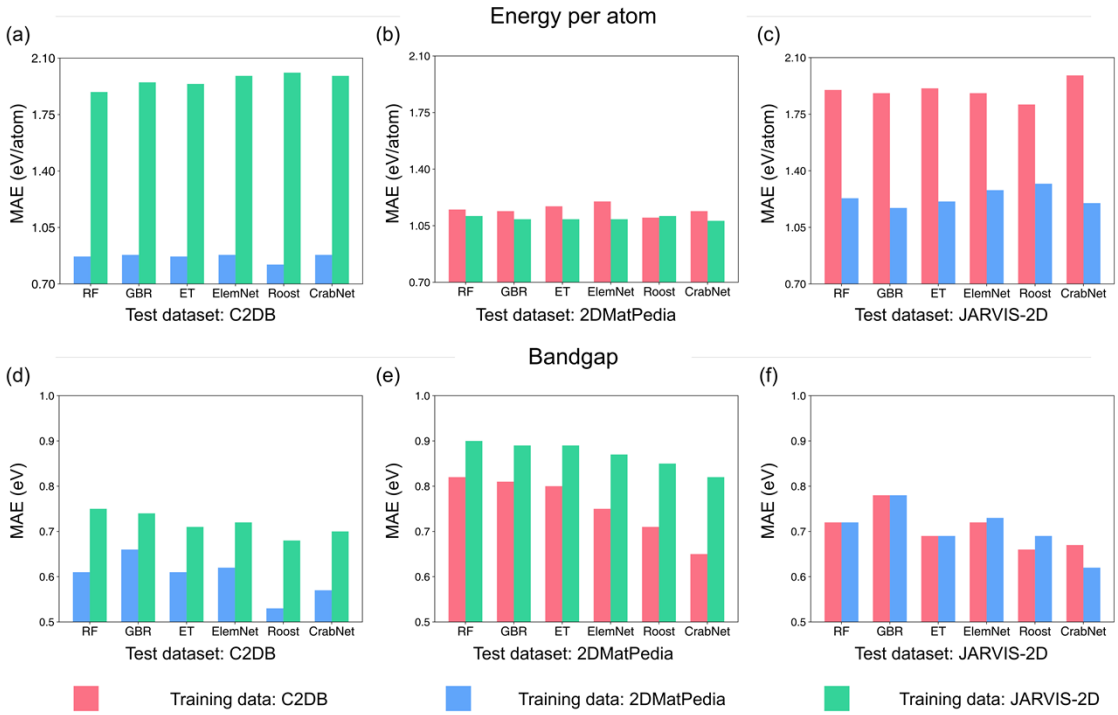

Figure S4. Database-to-database MAE performance of composition-based models for (a)–(c) energy per atom and (d)–(f) bandgap. Models were trained on two databases and tested on the remaining one: (a) and (d) test on C2DB, (b) and (e) test on 2DMatPedia, and (c) and (f) test on JARVIS-2D. Red markers indicate models trained on C2DB, blue markers indicate models trained on 2DMatPedia, and green markers indicate models trained on JARVIS-2D.

S8. Predictive performance of structure-based surrogate models

Table S10. MAE (unitless) of structure-based surrogate models for energy per atom and bandgap under database-to-database transfer settings. Results are reported for GCN, MEGNet, SchNet, CGCNN, and DeeperGATGNN across all database pairs (C2DB, 2DMatPedia, and JARVIS-2D).

Surrogate model	Property	C2DB to 2DMatPedia	C2DB to JARVIS-2D	2DMatPedia to C2DB	2DMatPedia to JARVIS-2D	JARVIS-2D to C2DB	JARVIS-2D to 2DMatPedia
GCN	Energy per atom (eV/atom)	1.15	1.85	0.89	1.13	1.96	1.06
MEGNet		1.16	1.88	0.88	1.17	1.89	1.09
SchNet		1.18	1.85	0.89	1.15	2.03	1.13
CGCNN		1.17	1.87	0.88	1.18	1.95	1.02
DeeperGATGNN		1.15	1.86	0.88	1.16	1.97	1.05
GCN	Bandgap (eV)	0.84	0.77	0.63	0.76	0.74	0.95
MEGNet		0.83	0.72	0.54	0.76	0.71	0.95
SchNet		0.80	0.78	0.55	0.73	0.72	1.00
CGCNN		0.77	0.71	0.56	0.75	0.70	0.99
DeeperGATGNN		0.77	0.69	0.58	0.78	0.70	0.94

Table S11. Spearman correlation of structure-based surrogate models for energy per atom and bandgap under database-to-database transfer settings. Results are reported for GCN, MEGNet, SchNet, CGCNN, and DeeperGATGNN across all database pairs (C2DB, 2DMatPedia, and JARVIS-2D).

Surrogate model	Property	C2DB to 2DMatPedia	C2DB to JARVIS-2D	2DMatPedia to C2DB	2DMatPedia to JARVIS-2D	JARVIS-2D to C2DB	JARVIS-2D to 2DMatPedia
GCN	Energy per atom (eV/atom)	0.84	0.86	0.91	0.94	0.89	0.89
MEGNet		0.84	0.86	0.92	0.96	0.85	0.87
SchNet		0.83	0.85	0.92	0.95	0.91	0.89
CGCNN		0.84	0.85	0.92	0.96	0.91	0.90
DeeperGATGNN		0.84	0.85	0.92	0.97	0.92	0.92
GCN	Bandgap (eV)	0.50	0.51	0.57	0.48	0.53	0.47
MEGNet		0.51	0.59	0.62	0.55	0.56	0.47
SchNet		0.53	0.53	0.60	0.55	0.56	0.47
CGCNN		0.57	0.63	0.62	0.51	0.57	0.47
DeeperGATGNN		0.53	0.60	0.60	0.48	0.61	0.52

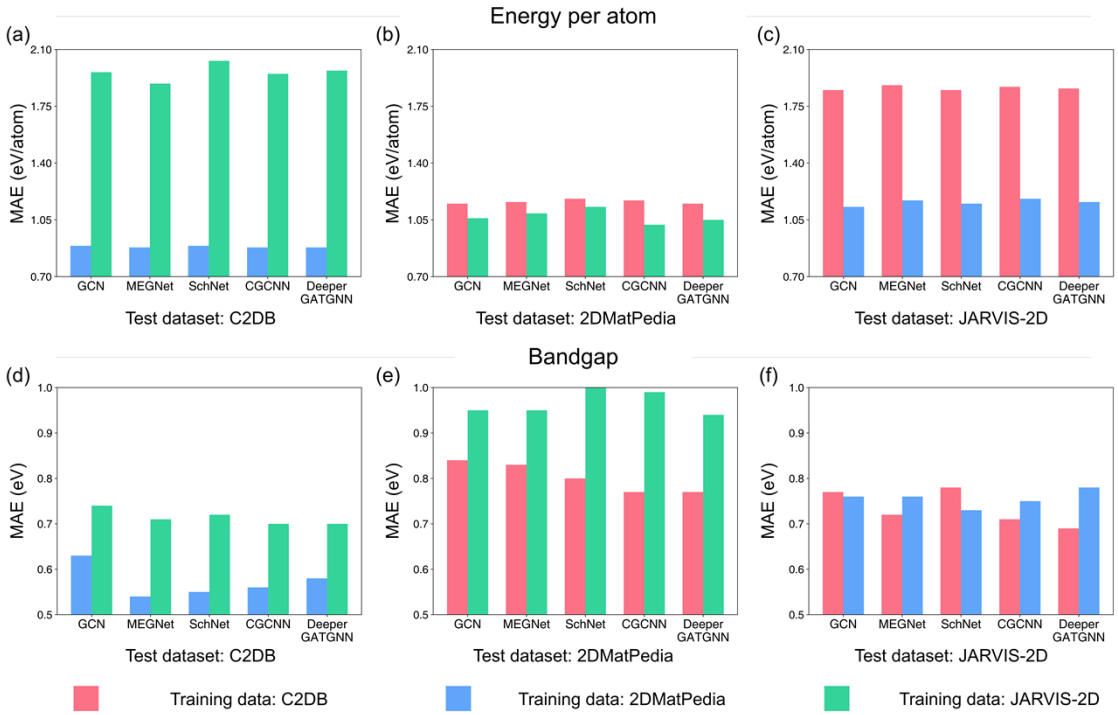

Figure S5. Database-to-database MAE performance of structure-based models for (a)-(c) energy per atom and (d)-(f) bandgap. Models were trained on two databases and tested on the remaining one: (a) and (d) test on C2DB, (b) and (e) test on 2DMatPedia, and (c) and (f) test on JARVIS-2D. Red markers indicate models trained on C2DB, blue markers indicate models trained on 2DMatPedia, and green markers indicate models trained on JARVIS-2D.

S9. Comparative analysis

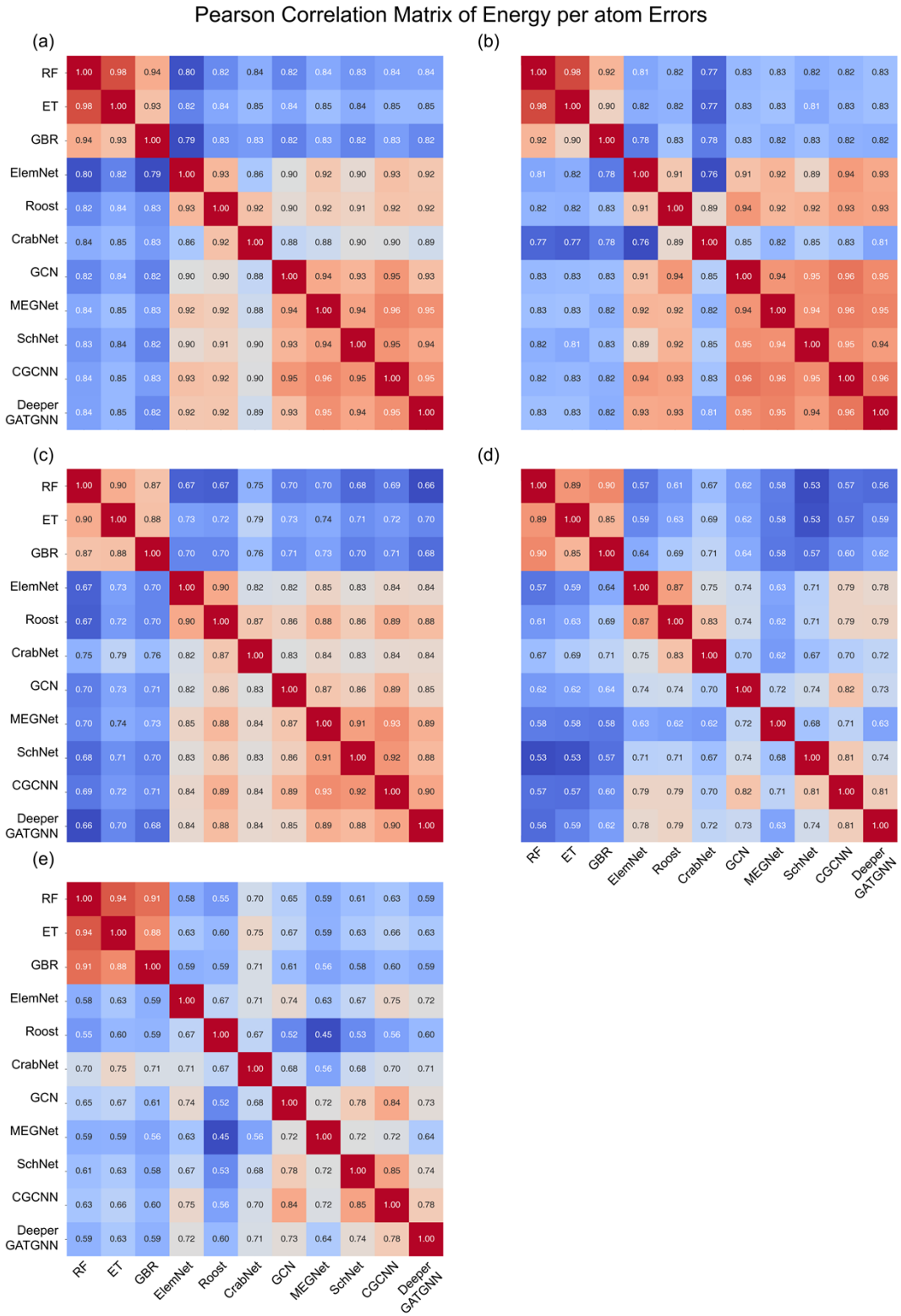

Figure S6. Pearson correlation matrix of prediction errors across all surrogate models for (a) C2DB to 2DMatPedia, (b) C2DB to JARVIS-2D, (c) 2DMatPedia to C2DB, (d) JARVIS-2D to C2DB, and (e) JARVIS-2D to 2DMatPedia in energy per atom predictions.

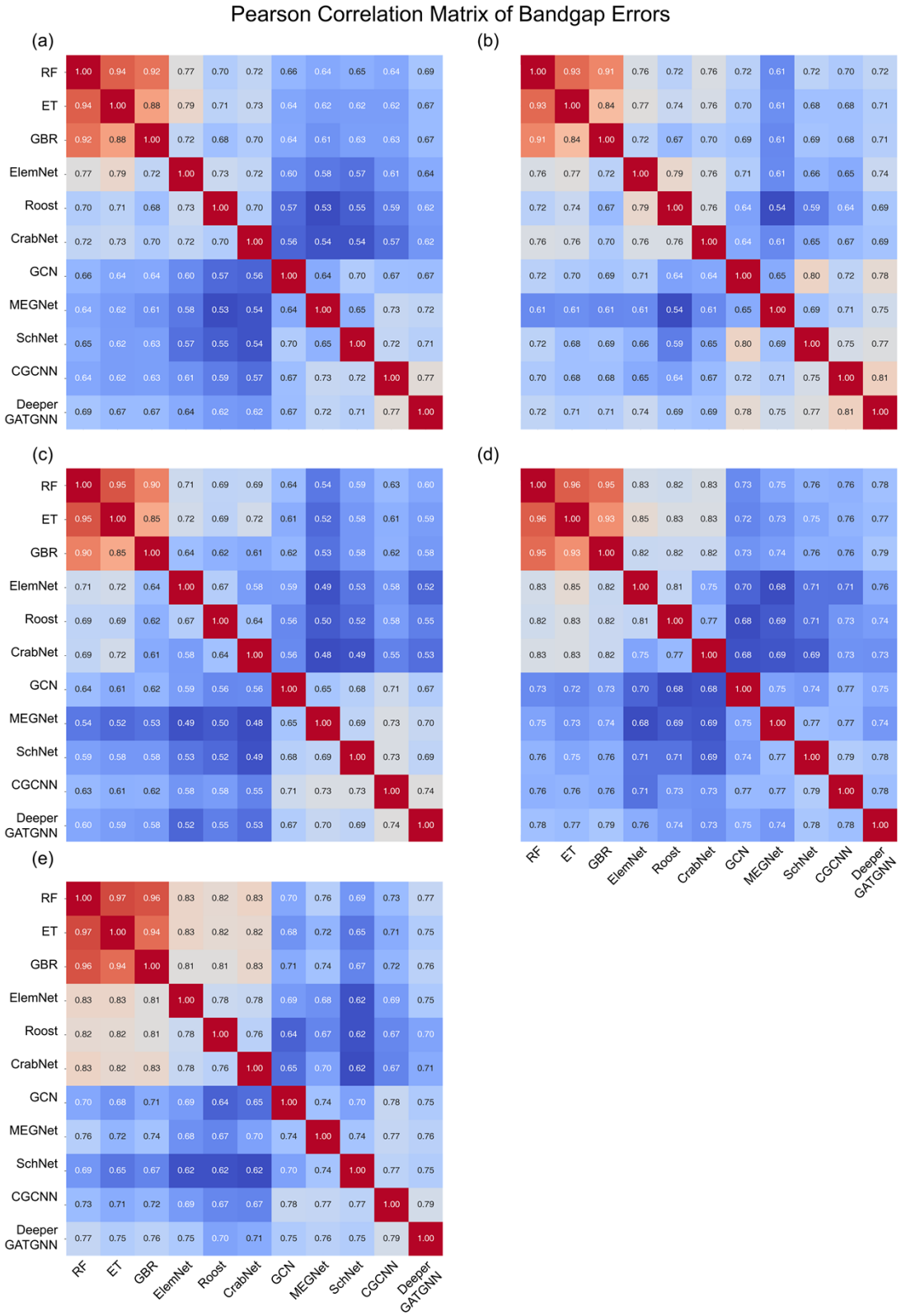

Figure S7. Pearson correlation matrix of prediction errors across all surrogate models for (a) C2DB to 2DMatPedia, (b) 2DMatPedia to C2DB, (c) 2DMatPedia to JARVIS-2D, (d) JARVIS-2D to C2DB, and (e) JARVIS-2D to 2DMatPedia in bandgap predictions.

References

1 K. Choudhary and F. Tavazza, *Comput Mater Sci*, 2019, **161**, 300–308.

2 Z. G. Yu and Y.-W. Zhang, *J Appl Phys*, 2015, **118**, 165706.

3 S. S. Yadavalli, G. Jones and M. Stamatakis, *Physical Chemistry Chemical Physics*, 2021, **23**, 15601–15612.

4 L. Breiman, *Random Forests*, 2001, vol. 45.

5 J. H. Friedman, *Greedy Function Approximation: A Gradient Boosting Machine*, 2001, vol. 29.

6 P. Geurts, D. Ernst and L. Wehenkel, *Mach Learn*, 2006, **63**, 3–42.

7 D. Jha, L. Ward, A. Paul, W. keng Liao, A. Choudhary, C. Wolverton and A. Agrawal, *Sci Rep*, DOI:10.1038/s41598-018-35934-y.

8 R. E. A. Goodall and A. A. Lee, *Nat Commun*, DOI:10.1038/s41467-020-19964-7.

9 A. Y. T. Wang, S. K. Kauwe, R. J. Murdock and T. D. Sparks, *NPJ Comput Mater*, DOI:10.1038/s41524-021-00545-1.

10 T. N. Kipf and M. Welling, .

11 C. Chen, W. Ye, Y. Zuo, C. Zheng and S. P. Ong, *Chemistry of Materials*, 2019, **31**, 3564–3572.

12 K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko and K.-R. Müller, .

13 T. Xie and J. C. Grossman, *Phys Rev Lett*, DOI:10.1103/PhysRevLett.120.145301.

14 S. S. Omee, S. Y. Louis, N. Fu, L. Wei, S. Dey, R. Dong, Q. Li and J. Hu, *Patterns*, DOI:10.1016/j.patter.2022.100491.