Supplementary Information

Contents

S1 Experimental dataset evaluation 1

S2 Category based Leave-one-material-out split 1

S3 Analysis of feature space 2

S4 Experimental setup 2

S5 Experimental results on random splits 3

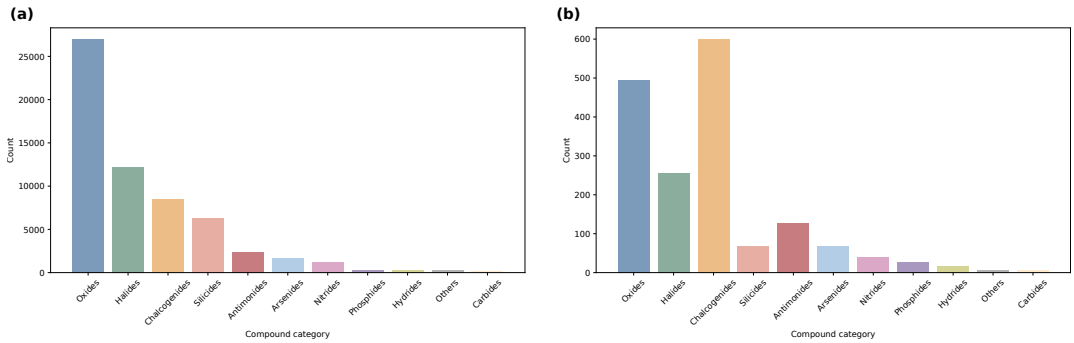
S6 Experimental results on domain-based splits 3

S7 Classical machine learning baselines on learned representations 14

S8 Experimental results on atomic encoding SHAP values of SVR 14

S1 Experimental dataset evaluation

Supplementary Figure 1 is a supplement to Fig. 1b–d. It further shows the distributions of computational PBE and experimental bandgap data across material categories used in this paper. The comparison indicates that the experimental dataset is dominated by chalcogenides, which have been actively investigated as wide-bandgap semiconductors due to their tunable electronic properties and relevance to optoelectronic applications. As a result, models trained primarily on computational data may learn biased representations when applied without sufficient experimental supervision. The data are categorized according to the standard in Supplementary Table 1.



Supplementary Figure 1. a, Number of materials in the computational dataset for each material category. **b,** Number of materials in the experimental dataset for each material category.

Supplementary Table 1. Categorization of materials based on chemical composition.

Category	Arsenides	Antimonides	Silicides	Halides	Chalcogenides	Oxides
Key element(s)	As	Sb	Si	F, Cl, Br, I	S, Se, Te	O
Category	Nitrides	Phosphides	Carbides	Hydrides	Others	
Key element(s)	N	P	C	H	Not classified above	

S2 Category based Leave-one-material-out split

The LOMO split uses the categorization scheme defined in Supplementary Table 1. Compounds that belong to multiple categories are removed to ensure that each split is mutually exclusive and to prevent distribution leakage. As shown in Supplementary Table 2, only 1254 crystals are retained for the LOMO setting. Supplementary Table 3 further summarizes the number of compounds in each category.

Supplementary Table 2. Category distribution showing the number of categories each compound belongs to, including those that do not belong to any category.

Category #	1	2	3	Not belong to any
Count	1254	406	38	7

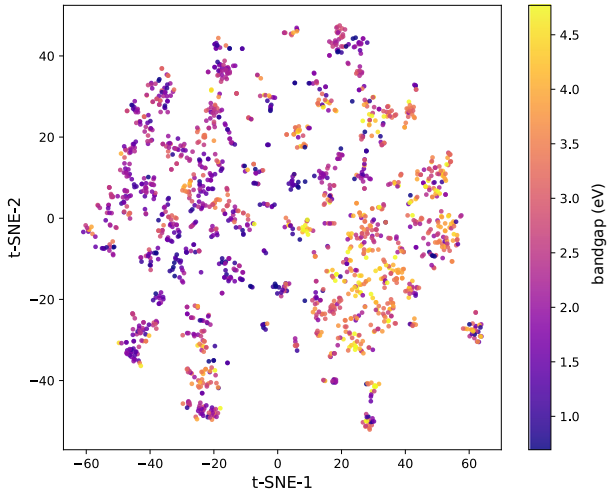
Supplementary Table 3. Number of compounds in each category.

Category	Chalcogenides	Oxides	Halides	Nitrides	Phosphides
Count	486	481	159	36	27
Category	Antimonides	Hydrides	Arsenides	Silicides	Carbides
Count	18	16	16	8	7

S3 Analysis of feature space

In this paper, we adopt the fixed-dimensional material representations extracted from CGCNN as the feature space. To examine how the target property is organized under this representation, we analyze the distribution of experimental bandgap values in the corresponding t-SNE embedding. From Supplementary Figure 2, we observe that the bandgap values vary smoothly across the embedded space, with higher bandgaps predominantly located on one side of the manifold and lower bandgaps on the opposite side.

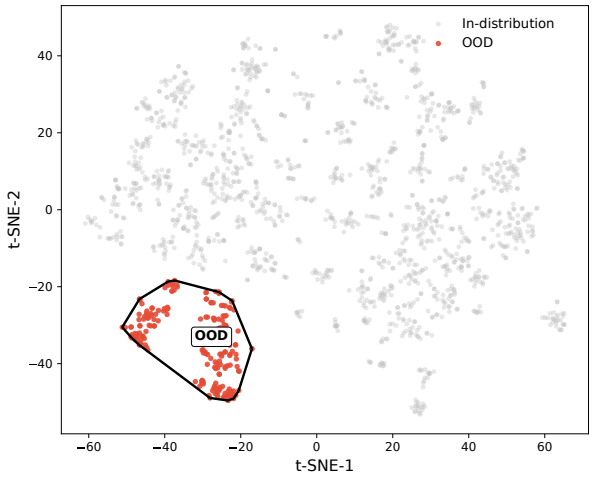
Although the representation method employed does not explicitly encode detailed geometric or topological information, it captures the statistical frequency of different atom species and coarse-grained interatomic distance information. These factors are well known to play a dominant role in determining the overall electronic properties of materials and can significantly limit the range of bandgap values.



Supplementary Figure 2. t-SNE visualization of the CGCNN-derived feature representation, with points colored by bandgap values.

S4 Experimental setup

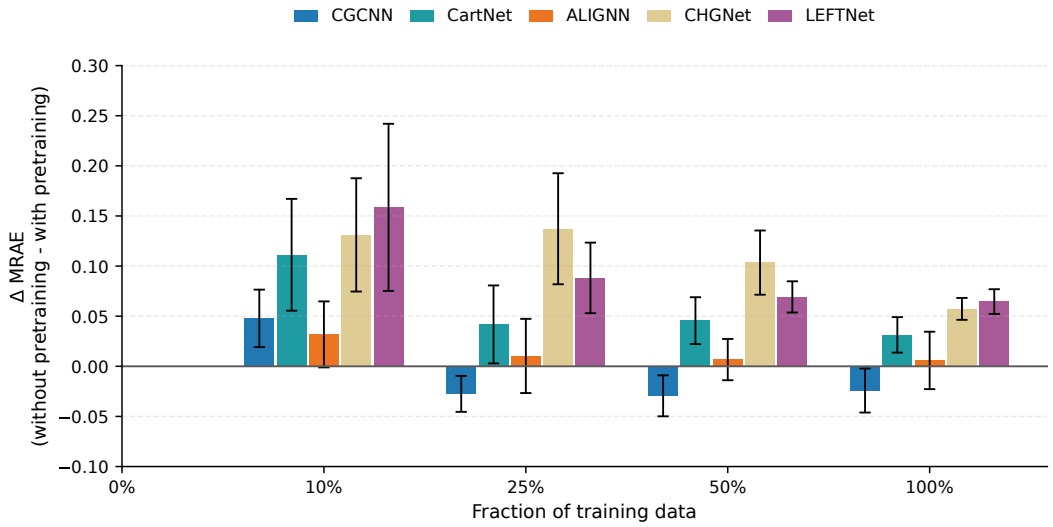
All deep models were trained using stochastic gradient descent (SGD) with a learning rate of 10^{-3}, momentum of 0.9, and a batch size of 64. A fixed random seed (42) was used for reproducibility. Model checkpointing was performed based on the lowest validation MRE. Pretraining runs used 200 epochs, while all downstream tasks were trained for 100 epochs. For each data split or fold, models were re-initialized and trained independently on a single GPU when available. The architecture and configurations of the GNN models are summarised in Supplementary Table 4. Hyperparameter tuning is applied to the classical machine learning models, with details provided in Supplementary Table 5.



Supplementary Figure 3. *t*-SNE projection of crystal graph embeddings from the experimental dataset. An isolated cluster of materials, automatically selected using density-based clustering, is highlighted in red and designated as out-of-distribution (OOD), while the remaining samples form the in-distribution region (grey).

S5 Experimental results on random splits

To more directly quantify the impact of computational pretraining, Supplementary Figure 4 presents the difference in MRAE between models trained without and with pretraining across different fractions of the experimental training data. The metric is defined as ΔMRAE = MRAE (without pretraining) − MRAE (with pretraining), such that positive values indicate performance improvements from pretraining, whereas negative values indicate performance degradation. This representation provides a complementary view to Fig. 2, highlighting how the benefit of pretraining varies across architectures and training data regimes.



Supplementary Figure 4. Effect of pretraining across different data fractions. Difference in MRAE between models trained without and with computational pretraining across different experimental data fractions. Positive values indicate improved performance from pretraining. Error bars denote propagated standard deviation.

S6 Experimental results on domain-based splits

Across the LOMO split (Supplementary Table 6, Supplementary Table 7 and Supplementary Table 8), silicides and antimonides show larger prediction errors than others for nearly all architectures and metrics. These indicate poor generalization when these

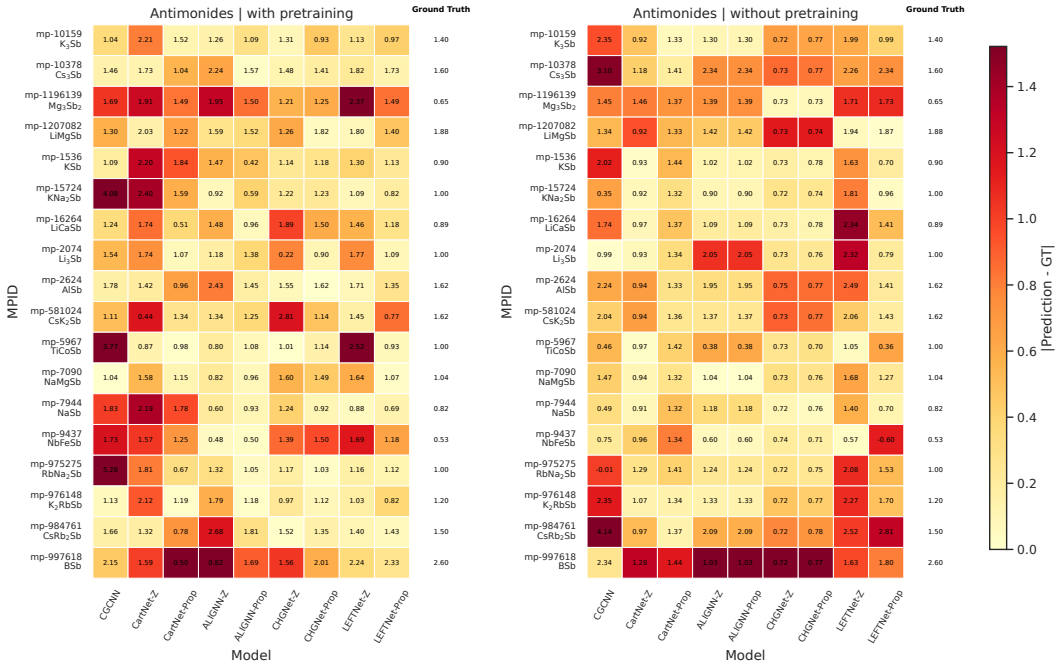
Supplementary Table 4. Key architectural and graph construction configurations of the GNN models used in this study.

Model	Configuration	Value
CGCNN	Atom embedding dimension	64
	Hidden dimension	128
	Number of convolution layers	3
	Number of output layers	1
ALIGNN	Hidden dimension	128
	Number of ALIGNN layers	4
	Number of GCN layers	4
	Number of Gaussian basis functions	40
	Atom cutoff radius (Å)	6.0
	Bond cutoff radius (Å)	3.0
CHGNet	Atom feature dimension	64
	Bond feature dimension	64
	Angle feature dimension	64
	Number of convolution layers	4
	Radial basis size	31
	Angular basis size	31
LEFTNet	Hidden channels	128
	Number of layers	4
CartNet	Input dimension	256
	Radial basis dimension	64
	Number of layers	4
	Maximum neighbors	12
	Cutoff radius (Å)	8.0

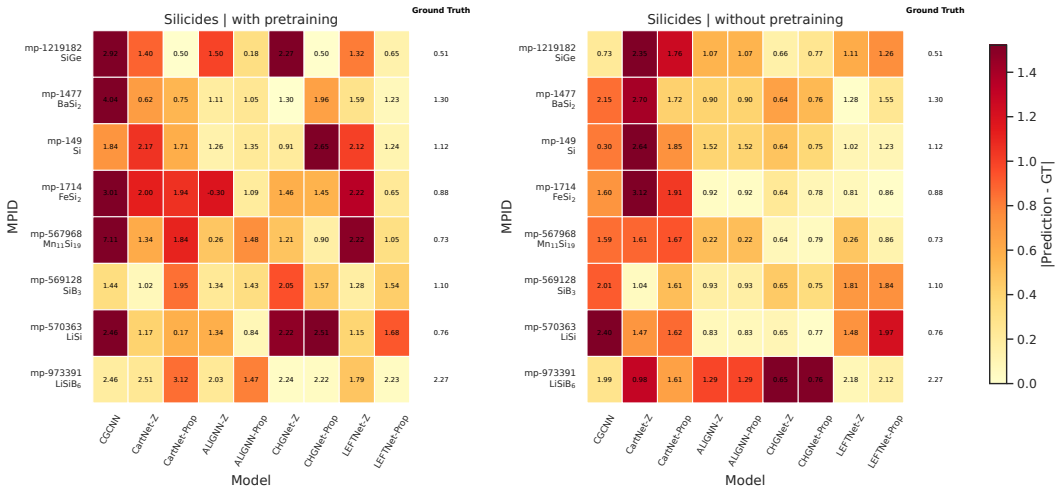
material families are excluded from training. This trend is further illustrated by the error heatmaps in Supplementary Figure 5 and Supplementary Figure 6, which visualize prediction errors for the two most challenging categories.

The results also reveal architecture-dependent robustness under category-level distribution shift. For instance, ALIGNN maintains relatively stable MRAE across most categories in the pretraining setting (generally $\sim$0.3–0.5), whereas CGCNN and LEFTNet show larger performance variance across chemical families.

Pretraining does not consistently improve performance under LOMO split. In the difficult categories (silicides and antimonides), pretraining sometimes increases error. For example, CGCNN MRAE on silicides increases from 0.867 (without pretraining) to 2.661 (with pretraining), and LEFTNet shows similarly large silicide errors in both settings.



Supplementary Figure 5. Heatmap of bandgap prediction errors for the antimonides split. The left and right panels correspond to models with and without computational pretraining, respectively. Color indicates the absolute error $|\text{Prediction} - \text{GT}|$, where GT denotes the ground-truth bandgap. Each cell is annotated with the predicted bandgap, and the rightmost column lists the ground-truth bandgap for each Materials Project ID (MPID).



Supplementary Figure 6. Heatmap of bandgap prediction errors for the silicides split. The left and right panels correspond to models with and without computational pretraining, respectively. Color indicates the absolute error $|\text{Prediction} - \text{GT}|$, where GT denotes the ground-truth bandgap. Each cell is annotated with the predicted bandgap, and the rightmost column lists the ground-truth bandgap for each Materials Project ID (MPID).

Supplementary Table 5. Hyperparameter search space for classical machine learning models on the random split task. Bold values indicate the configuration selected based on the lowest validation MRAE and used for reporting test performance. Linear regression has no tunable hyperparameters under our setup and is therefore omitted from this table.

Model	Hyperparameter	Candidate values
SVR	Regularization C	1, **10**, 100
	RBF kernel width γ	0.01, **0.1**, 1
	ε-insensitive loss	0.01, **0.1**, 0.5
	Kernel type	**rbf**, linear, poly, sigmoid
Random Forest	Number of trees	200, 400, **800**
	Maximum depth	**None**, 10, 20
	Min. samples for split	**2**, 4, 8
	Min. samples per leaf	**1**, 2, 4

Supplementary Table 6. MRAE scores of each model on the LOMO split. For each category, that category is left out as the test set. Standard deviation is not reported because this split uses a single fold. All values are reported to three decimal places.

With pretraining

Model	Antimonides	Arsenides	Carbides	Chalcogenides	Halides	Hydrides	Nitrides	Oxides	Phosphides	Silicides
CGCNN	0.988	0.502	0.590	0.329	0.391	0.357	0.447	0.493	0.375	2.661
CartNet	0.716	0.401	0.250	0.318	0.422	0.542	0.824	0.371	0.435	0.619
ALIGNN	0.448	0.358	0.190	0.312	0.367	0.307	0.496	0.332	0.324	0.664
CHGNet	0.443	0.470	0.236	0.279	0.410	0.355	0.444	0.287	0.369	0.972
LEFTNet	0.555	0.498	0.255	0.269	0.363	0.299	0.488	0.286	0.369	0.896

Without pretraining

Model	Antimonides	Arsenides	Carbides	Chalcogenides	Halides	Hydrides	Nitrides	Oxides	Phosphides	Silicides
CGCNN	0.680	0.546	0.214	0.397	0.405	0.375	0.535	0.342	0.430	0.867
CartNet	0.387	0.358	0.106	0.365	0.370	0.379	0.467	0.351	0.323	1.072
ALIGNN	0.350	0.289	0.207	0.342	0.449	0.392	0.408	0.314	0.376	0.400
CHGNet	0.376	0.426	0.157	0.372	0.385	0.427	0.468	0.344	0.388	0.360
LEFTNet	0.687	0.620	0.199	0.447	0.441	0.395	0.465	0.334	0.291	0.457
LR	0.805	0.708	0.373	0.409	0.485	1.013	0.561	0.354	0.268	4.084
RFR	0.792	0.484	0.243	0.366	0.440	0.453	0.717	0.350	0.371	1.523
SVR	0.632	0.629	0.252	0.399	0.468	0.427	0.505	0.321	0.295	1.759

Supplementary Table 7. MAE scores of each model on the LOMO split. For each category, that category is left out as the test set. Standard deviation is not reported because this split uses a single fold. All values are reported to three decimal places.

With pretraining

Model	Antimonides	Arsenides	Carbides	Chalcogenides	Halides	Hydrides	Nitrides	Oxides	Phosphides	Silicides
CGCNN	0.939	0.601	1.463	0.621	0.840	0.790	0.969	1.522	0.605	2.077
CartNet	0.768	0.517	0.635	0.601	0.879	0.966	1.295	1.040	0.637	0.573
ALIGNN	0.528	0.449	0.499	0.593	0.899	0.827	0.870	0.862	0.526	0.505
LEFTNet	0.472	0.567	0.631	0.511	0.710	0.609	0.939	0.728	0.515	0.748
CHGNet	0.456	0.505	0.624	0.503	0.756	0.606	0.753	0.703	0.560	0.685

Without pretraining

Model	Antimonides	Arsenides	Carbides	Chalcogenides	Halides	Hydrides	Nitrides	Oxides	Phosphides	Silicides
CGCNN	0.780	0.504	0.555	0.667	1.013	0.925	1.003	0.851	0.552	0.788
CartNet	0.477	0.430	0.329	0.726	0.968	1.071	1.092	1.003	0.461	0.941
ALIGNN	0.426	0.345	0.526	0.656	1.153	1.046	0.827	0.754	0.601	0.393
LEFTNet	0.748	0.616	0.537	0.988	1.205	1.073	1.003	0.850	0.435	0.349
CHGNet	0.540	0.415	0.480	0.644	0.932	1.239	1.181	0.945	0.598	0.474

Supplementary Table 8. R^2 scores of each model on the LOMO split. For each category, that category is left out as the test set. Standard deviation is not reported because this split uses a single fold. All values are reported to three decimal places.

With pretraining

Model	Antimonides	Arsenides	Carbides	Chalcogenides	Halides	Hydrides	Nitrides	Oxides	Phosphides	Silicides
CGCNN	-8.422	-1.559	-6.931	-0.063	-0.263	0.284	-0.122	-2.758	-0.122	-29.173
CartNet	-2.259	-1.617	-0.259	0.113	-0.441	-0.157	-2.432	-0.655	-0.129	-0.970
ALIGNN	-1.045	-0.637	0.026	0.041	-0.487	0.106	0.152	-0.205	0.098	-0.514
CHGNet	-0.452	-1.564	-0.269	0.319	-0.064	0.524	0.230	0.065	-0.034	-2.279
LEFTNet	-0.959	-1.207	-0.223	0.282	0.106	0.619	-0.149	0.072	0.018	-1.997

Without pretraining

Model	Antimonides	Arsenides	Carbides	Chalcogenides	Halides	Hydrides	Nitrides	Oxides	Phosphides	Silicides
CGCNN	-2.969	-1.651	0.000	-0.131	-0.756	0.002	-0.086	-0.163	0.028	-2.059
CartNet	-0.621	-0.363	0.379	-0.353	-0.749	-0.433	-0.567	-0.568	0.327	-3.952
ALIGNN	-0.414	0.237	-0.061	-0.169	-1.223	-0.224	0.081	0.044	0.011	0.078
CHGNet	-1.083	-0.260	-0.278	-0.072	-0.516	-0.720	-0.655	-0.424	-0.081	-0.742
LEFTNet	-2.017	-1.821	-0.436	-1.314	-1.547	-0.308	-0.138	-0.172	0.327	0.206
LR	-2.867	-2.823	-1.118	0.062	-1.106	-8.971	-0.759	0.065	0.274	-55.401
RFR	-1.998	-0.927	-0.080	0.050	-0.383	0.052	-0.162	-0.622	0.240	-5.759
SVR	-2.079	-1.914	-0.304	-0.551	-1.341	0.044	0.078	-0.054	0.275	-8.261

Supplementary Table 9. Average MRAE under the LOMO split by category, comparing PBE computational errors with GNN models (excluding classical machine learning models) trained with pretraining and without pretraining.

Category	PBE	Models with pretraining	Models without pretraining
Antimonides	1.358	0.590	0.496
Arsenides	1.157	0.450	0.448
Carbides	0.250	0.321	0.177
Chalcogenides	0.502	0.301	0.384
Halides	0.440	0.398	0.410
Hydrides	0.484	0.389	0.394
Nitrides	0.633	0.542	0.469
Oxides	0.340	0.355	0.337
Phosphides	0.793	0.388	0.362
Silicides	1.770	1.106	0.631

Supplementary Table 10. Model performance under different evaluation splits, reported as mean ± standard deviation. Metrics include mean relative absolute error (MRAE), mean absolute error (MAE), and coefficient of determination (R^2). Results are shown for models with and without computational pretraining where applicable. This is a supplement of Fig. 3c-f.

Split	Model	MRAE	MAE	R^2
Random split with pretraining	CGCNN	0.295$_{(0.014)}$	0.603$_{(0.038)}$	0.235$_{(0.068)}$
	CartNet	0.306$_{(0.009)}$	0.606$_{(0.042)}$	0.273$_{(0.062)}$
	ALIGNN	0.280$_{(0.013)}$	0.550$_{(0.030)}$	0.407$_{(0.051)}$
	CHGNet	**0.242**$_{(0.004)}$	0.486$_{(0.005)}$	0.487$_{(0.014)}$
	LEFTNet	0.242$_{(0.003)}$	**0.485**$_{(0.007)}$	**0.507**$_{(0.017)}$
Random split without pretraining	CGCNN	0.271$_{(0.017)}$	0.576$_{(0.019)}$	0.313$_{(0.051)}$
	CartNet	0.337$_{(0.015)}$	0.738$_{(0.048)}$	-0.342$_{(1.030)}$
	ALIGNN	0.286$_{(0.025)}$	0.586$_{(0.054)}$	0.369$_{(0.090)}$
	CHGNet	0.299$_{(0.010)}$	0.613$_{(0.022)}$	0.298$_{(0.056)}$
	LEFTNet	0.307$_{(0.012)}$	0.616$_{(0.020)}$	0.246$_{(0.130)}$
	LR	0.332$_{(0.005)}$	0.637$_{(0.014)}$	0.306$_{(0.008)}$
	RFR	0.284$_{(0.005)}$	0.535$_{(0.007)}$	0.453$_{(0.016)}$
	SVR	**0.261**$_{(0.007)}$	**0.510**$_{(0.013)}$	**0.497**$_{(0.018)}$
Chemical system with pretraining	CGCNN	**0.275**$_{(0.029)}$	0.562$_{(0.044)}$	0.411$_{(0.148)}$
	CartNet	0.325$_{(0.051)}$	0.617$_{(0.076)}$	0.245$_{(0.285)}$
	ALIGNN	0.287$_{(0.033)}$	0.576$_{(0.063)}$	0.412$_{(0.118)}$
	CHGNet	0.287$_{(0.041)}$	0.531$_{(0.054)}$	0.457$_{(0.146)}$
	LEFTNet	0.283$_{(0.042)}$	**0.529**$_{(0.064)}$	**0.484**$_{(0.135)}$
Chemical system without pretraining	CGCNN	0.302$_{(0.040)}$	0.601$_{(0.074)}$	0.345$_{(0.137)}$
	CartNet	0.328$_{(0.034)}$	0.687$_{(0.069)}$	0.183$_{(0.271)}$
	ALIGNN	0.295$_{(0.045)}$	0.594$_{(0.080)}$	0.398$_{(0.125)}$
	CHGNet	0.366$_{(0.037)}$	0.711$_{(0.135)}$	0.197$_{(0.306)}$
	LEFTNet	0.331$_{(0.036)}$	0.614$_{(0.037)}$	0.372$_{(0.080)}$
	LR	0.361$_{(0.033)}$	0.660$_{(0.047)}$	0.328$_{(0.094)}$
	RFR	0.303$_{(0.044)}$	0.550$_{(0.061)}$	**0.494**$_{(0.105)}$
	SVR	**0.294**$_{(0.046)}$	**0.543**$_{(0.059)}$	0.477$_{(0.104)}$
Periodic groups with pretraining	CGCNN	0.326$_{(0.074)}$	0.620$_{(0.096)}$	0.287$_{(0.192)}$
	CartNet	0.345$_{(0.054)}$	0.669$_{(0.100)}$	-0.169$_{(0.795)}$
	ALIGNN	0.334$_{(0.061)}$	0.640$_{(0.100)}$	0.263$_{(0.193)}$
	CHGNet	**0.308**$_{(0.058)}$	**0.566**$_{(0.083)}$	0.346$_{(0.258)}$
	LEFTNet	0.324$_{(0.097)}$	0.579$_{(0.089)}$	**0.377**$_{(0.174)}$
Periodic groups without pretraining	CGCNN	0.378$_{(0.077)}$	0.716$_{(0.083)}$	0.096$_{(0.164)}$
	CartNet	0.382$_{(0.083)}$	0.757$_{(0.092)}$	-0.035$_{(0.289)}$
	ALIGNN	**0.342**$_{(0.052)}$	**0.679**$_{(0.103)}$	0.179$_{(0.251)}$
	CHGNet	0.370$_{(0.052)}$	0.764$_{(0.131)}$	-0.057$_{(0.387)}$
	LEFTNet	0.448$_{(0.183)}$	0.832$_{(0.191)}$	-0.223$_{(0.490)}$
	LR	0.481$_{(0.113)}$	0.913$_{(0.261)}$	-0.582$_{(0.885)}$
	RFR	0.386$_{(0.103)}$	0.683$_{(0.096)}$	**0.228**$_{(0.173)}$
	SVR	0.400$_{(0.097)}$	0.773$_{(0.265)}$	-0.051$_{(0.628)}$
Crystal system with pretraining	CGCNN	0.319$_{(0.084)}$	0.619$_{(0.130)}$	0.208$_{(0.402)}$
	CartNet	0.343$_{(0.071)}$	0.672$_{(0.090)}$	-0.104$_{(1.013)}$
	ALIGNN	0.298$_{(0.044)}$	0.579$_{(0.086)}$	0.431$_{(0.140)}$
	CHGNet	**0.282**$_{(0.046)}$	**0.538**$_{(0.079)}$	**0.483**$_{(0.128)}$
	LEFTNet	0.285$_{(0.053)}$	0.557$_{(0.088)}$	0.476$_{(0.105)}$
Crystal system without pretraining	CGCNN	0.319$_{(0.046)}$	0.617$_{(0.098)}$	0.356$_{(0.163)}$
	CartNet	0.333$_{(0.065)}$	0.746$_{(0.205)}$	0.019$_{(0.482)}$
	ALIGNN	**0.304**$_{(0.039)}$	0.643$_{(0.125)}$	0.330$_{(0.213)}$
	CHGNet	0.375$_{(0.064)}$	0.708$_{(0.144)}$	0.237$_{(0.320)}$
	LEFTNet	0.344$_{(0.046)}$	0.648$_{(0.081)}$	0.338$_{(0.151)}$
	LR	0.379$_{(0.076)}$	0.701$_{(0.130)}$	0.236$_{(0.286)}$
	RFR	0.318$_{(0.049)}$	0.587$_{(0.087)}$	**0.465**$_{(0.085)}$
	SVR	0.307$_{(0.044)}$	**0.575**$_{(0.080)}$	0.444$_{(0.123)}$

Supplementary Table 11. Pretrained model performance on the experimental test set (best in bold, second best underlined).

Model	MAE ↓	MRAE ↓	R^2 ↑
ALIGNN	<u>0.811</u>	<u>0.346</u>	-0.146
CartNet	0.913	0.395	<u>-0.308</u>
CGCNN	**0.627**	**0.295**	**0.271**
CHGNet	1.456	0.589	-2.531
LEFTNet	1.165	0.489	-0.949

Supplementary Table 12. Performance of different models on the validation and test sets, reported as mean ± standard deviation (**best in bold**, <u>second best</u> in underline). Higher values indicate better performance (↑), otherwise (↓).

Model	Validation set			Test set		
	MAE(eV) ↓	MRAE ↓	R^2 ↑	MAE(eV) ↓	MRAE ↓	R^2 ↑
Without pretraining (fine-tuning only)						
CGCNN	$0.566_{(0.044)}$	$\mathbf{0.277_{(0.030)}}$	$0.425_{(0.092)}$	$0.576_{(0.019)}$	<u>$0.271_{(0.017)}$</u>	$0.313_{(0.051)}$
CartNet-Z	$0.660_{(0.092)}$	$0.322_{(0.048)}$	$0.214_{(0.221)}$	$0.738_{(0.048)}$	$0.337_{(0.015)}$	$-0.342_{(1.030)}$
CartNet-Prop	$0.655_{(0.109)}$	$0.315_{(0.057)}$	$0.054_{(0.656)}$	$0.703_{(0.067)}$	$0.312_{(0.023)}$	$-0.056_{(0.434)}$
ALIGNN-Z	$0.598_{(0.056)}$	$0.299_{(0.036)}$	$0.401_{(0.091)}$	$0.589_{(0.041)}$	$0.279_{(0.017)}$	$0.368_{(0.073)}$
ALIGNN-Prop	$0.573_{(0.073)}$	<u>$0.288_{(0.051)}$</u>	$0.421_{(0.117)}$	$0.586_{(0.054)}$	$0.286_{(0.025)}$	$0.369_{(0.090)}$
CHGNET-Z	$0.634_{(0.055)}$	$0.333_{(0.038)}$	$0.300_{(0.135)}$	$0.613_{(0.022)}$	$0.299_{(0.010)}$	$0.298_{(0.056)}$
CHGNET-Prop	$0.635_{(0.061)}$	$0.319_{(0.040)}$	$0.325_{(0.081)}$	$0.651_{(0.042)}$	$0.316_{(0.012)}$	$0.214_{(0.091)}$
LEFTNet-Z	$0.614_{(0.051)}$	$0.322_{(0.043)}$	$0.361_{(0.103)}$	$0.616_{(0.020)}$	$0.307_{(0.012)}$	$0.246_{(0.130)}$
LEFTNet-Prop	$0.600_{(0.104)}$	$0.304_{(0.044)}$	$0.380_{(0.183)}$	$0.612_{(0.126)}$	$0.297_{(0.029)}$	$0.256_{(0.302)}$
LR	$0.650_{(0.055)}$	$0.356_{(0.049)}$	$0.357_{(0.067)}$	$0.637_{(0.014)}$	$0.332_{(0.005)}$	$0.306_{(0.008)}$
RFR	$\mathbf{0.541_{(0.047)}}$	$0.296_{(0.037)}$	$\mathbf{0.510_{(0.054)}}$	<u>$0.535_{(0.007)}$</u>	$0.284_{(0.005)}$	<u>$0.453_{(0.016)}$</u>
SVR	<u>$0.544_{(0.065)}$</u>	$0.293_{(0.051)}$	<u>$0.480_{(0.098)}$</u>	$\mathbf{0.510_{(0.013)}}$	$\mathbf{0.261_{(0.007)}}$	$\mathbf{0.497_{(0.018)}}$
With pretraining						
CGCNN	$0.621_{(0.070)}$	$0.322_{(0.050)}$	$0.227_{(0.286)}$	$0.603_{(0.038)}$	$0.295_{(0.014)}$	$0.235_{(0.068)}$
CartNet-Z	$0.613_{(0.057)}$	$0.323_{(0.033)}$	$0.264_{(0.234)}$	$0.606_{(0.042)}$	$0.306_{(0.009)}$	$0.273_{(0.062)}$
CartNet-Prop	$0.607_{(0.079)}$	$0.314_{(0.040)}$	$0.154_{(0.506)}$	$0.601_{(0.019)}$	$0.289_{(0.007)}$	$0.147_{(0.222)}$
ALIGNN-Z	$0.557_{(0.028)}$	$0.278_{(0.027)}$	$0.459_{(0.059)}$	$0.581_{(0.031)}$	$0.263_{(0.013)}$	$0.338_{(0.053)}$
ALIGNN-Prop	$0.529_{(0.047)}$	$\mathbf{0.266_{(0.032)}}$	$0.488_{(0.096)}$	$0.550_{(0.030)}$	$0.280_{(0.013)}$	$0.407_{(0.051)}$
CHGNET-Z	$0.535_{(0.053)}$	$0.289_{(0.039)}$	$0.472_{(0.105)}$	<u>$0.486_{(0.005)}$</u>	<u>$0.242_{(0.004)}$</u>	<u>$0.487_{(0.014)}$</u>
CHGNET-Prop	$\mathbf{0.516_{(0.038)}}$	$0.276_{(0.033)}$	<u>$0.506_{(0.091)}$</u>	$0.525_{(0.006)}$	$0.286_{(0.004)}$	$0.406_{(0.014)}$
LEFTNet-Z	<u>$0.520_{(0.044)}$</u>	$0.280_{(0.040)}$	$\mathbf{0.512_{(0.101)}}$	$\mathbf{0.485_{(0.007)}}$	$0.242_{(0.003)}$	$\mathbf{0.507_{(0.017)}}$
LEFTNet-Prop	$0.528_{(0.055)}$	<u>$0.268_{(0.044)}$</u>	$0.495_{(0.089)}$	$0.494_{(0.040)}$	$\mathbf{0.232_{(0.013)}}$	$0.484_{(0.061)}$

Supplementary Table 13. LOMO evaluation comparing Z encoding and atomic property embeddings with and without pretraining (MRAE).

With pretraining

Model	Encoding	Antimonides	Arsenides	Carbides	Chalcogenides	Halides	Hydrides	Nitrides	Oxides	Phosphides	Silicides
CartNet	Z	0.716	0.401	0.250	0.318	0.422	0.542	0.824	0.371	0.435	0.619
CartNet	Prop	0.267	0.414	0.322	0.297	0.423	0.508	0.520	0.497	0.319	1.008
ALIGNN	Z	0.448	0.358	0.190	0.312	0.367	0.307	0.496	0.332	0.324	0.664
ALIGNN	Prop	0.248	0.381	0.272	0.312	0.402	0.394	0.509	0.340	0.394	0.383
LEFTNet	Z	0.555	0.498	0.255	0.269	0.363	0.299	0.488	0.286	0.369	0.896
LEFTNet	Prop	0.309	0.348	0.271	0.275	0.334	0.222	0.520	0.294	0.334	0.346
CHGNet	Z	0.443	0.470	0.236	0.279	0.410	0.355	0.444	0.287	0.369	0.972
CHGNet	Prop	0.333	0.387	0.266	0.264	0.414	0.297	0.591	0.289	0.289	0.692

Without pretraining

Model	Encoding	Antimonides	Arsenides	Carbides	Chalcogenides	Halides	Hydrides	Nitrides	Oxides	Phosphides	Silicides
CartNet	Z	0.387	0.358	0.106	0.365	0.370	0.379	0.467	0.351	0.323	1.072
CartNet	Prop	0.435	0.498	0.121	0.359	0.387	0.488	0.421	0.361	0.244	0.907
ALIGNN	Z	0.350	0.289	0.207	0.342	0.449	0.392	0.408	0.314	0.376	0.400
ALIGNN	Prop	0.350	0.289	0.207	0.342	0.449	0.392	0.408	0.314	0.376	0.400
LEFTNet	Z	0.687	0.620	0.199	0.447	0.441	0.395	0.465	0.334	0.291	0.457
LEFTNet	Prop	0.500	0.438	0.199	0.327	0.507	0.364	0.453	0.325	0.333	0.540
CHGNet	Z	0.376	0.426	0.157	0.372	0.385	0.427	0.468	0.344	0.388	0.360
CHGNet	Prop	0.349	0.384	0.159	0.378	0.407	0.376	0.487	0.346	0.272	0.306

S7 Classical machine learning baselines on learned representations

We considered linear regression (LR), random forest regression (RFR) and support vector regression (SVR), representing linear, kernel-based, and tree-based learning paradigms, respectively. These models are used as probes to assess how much predictive signal is already encoded in the learned representation. We evaluated the models under the random split (Supplementary Figure 7a), feature-based split (Supplementary Figure 7b), periodic group split as a representative chemistry-based split (Supplementary Figure 7c), and crystal-system split as a structure-based split (Supplementary Figure 7d).

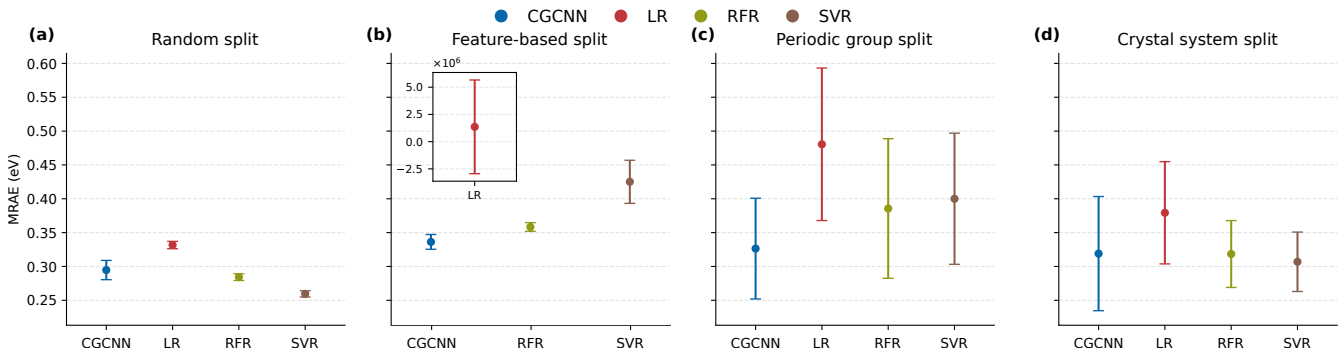
Across all evaluated tasks, we observed that nonlinear modeling (SVR and RFR) consistently outperformed LR. This indicates that linear mappings are insufficient to exploit the predictive information contained in the embedding fully. Under feature-based OOD, nonlinear models maintain strong performance, while the LR fails the task. This suggests that the predictive information in the embeddings retains local consistency in the held-out set, reflecting the role of local chemical and coordination motifs in determining electronic structure.

Among the nonlinear models, their relative strengths depend on the difficulty of the task. Under relatively easy settings, such as random and crystal-system splits, SVR achieves slightly better performance, likely benefiting from its smooth kernel-based interpolation. In contrast, under more challenging OOD scenarios, particularly feature-based and periodic-group splits, RFR exhibits stronger robustness, suggesting that segmentation-based models generalize more reliably when test samples exceed the training set.

At the same time, the performance of classical models, sometimes even rivaling or surpassing certain GNNs, should be interpreted with caution. These results do not directly indicate whether such models can capture material physics. Because the encoding is obtained at the compound level by averaging the atomic environment, the model acquires not an explicit crystal geometry, but rather an aggregated representation of structural information. Instead, these results indicate that the predictive information encoded in the learned representation is strongly correlated with the bandgap and can be effectively exploited for estimation by relatively simple models.

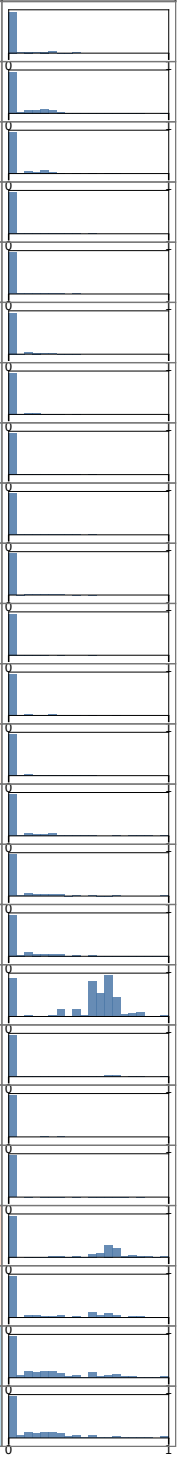
S8 Experimental results on atomic encoding SHAP values of SVR

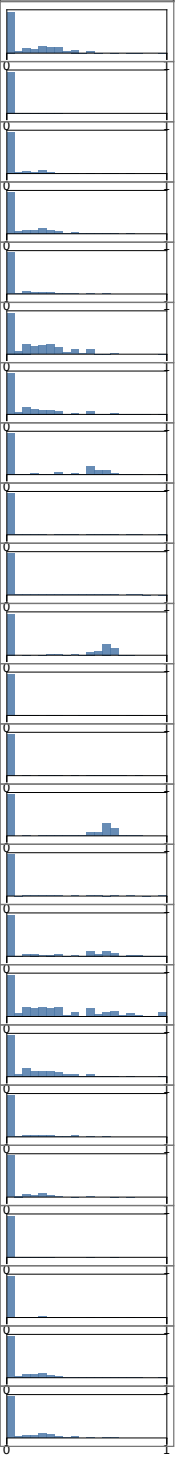
This section supplements Fig. 5 by providing the complete SHAP summary tables for the SVR model across all encoding features. The features are grouped into three segments: atom-level self features, neighbor-aggregated features, and radial basis function (RBF) distance features. For each feature, we report the mean SHAP value and the empirical distribution of the corresponding feature values in the training set. Feature distributions are normalized at the material level, where 0 indicates that none of the atoms in a given material falls into the corresponding bin and 1 indicates that all atoms fall into that bin.



Supplementary Figure 7. Test MRAE of CGCNN and classical machine learning baselines (LR, RFR, SVR) under different data splits: **a,** random split, **b,** feature-based OOD split (LR shown in the inset due to the difference in scale), **c,** periodic group split, and **d,** crystal system split. Error bars indicate one standard deviation across runs.

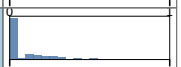
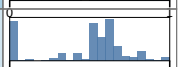
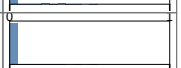
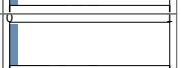
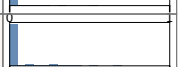
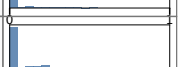
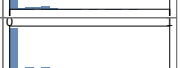
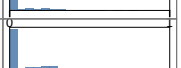
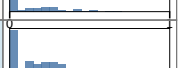
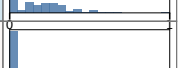
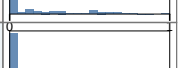
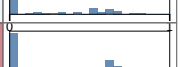
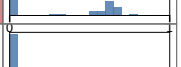
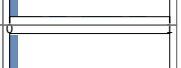
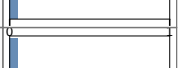
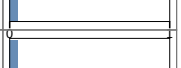
85

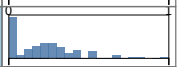
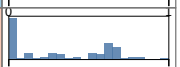
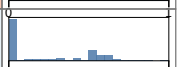
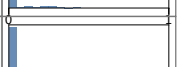
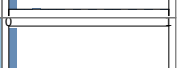
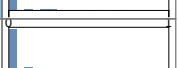
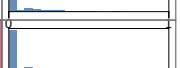
Idx	Property	mean SHAP	Distribution
1	Grp: 1	-0.00015	
2	Grp: 2	0.00063	
3	Grp: 3	0.00024	
4	Grp: 4	0.00026	
5	Grp: 5	-0.00010	
6	Grp: 6	-0.00038	
7	Grp: 7	0.00000	
8	Grp: 8	-0.00125	
9	Grp: 9	-0.00118	
10	Grp: 10	-0.00011	
11	Grp: 11	-0.00014	
12	Grp: 12	0.00060	
13	Grp: 13	0.00100	
14	Grp: 14	0.00000	
15	Grp: 15	-0.00004	
16	Grp: 16	0.00000	
17	Grp: 17	-0.00014	
18	Grp: 18	0.00065	
19	Grp: unk	0.00000	
20	Per: 1	0.00000	
21	Per: 2	0.00318	
22	Per: 3	0.00110	
23	Per: 4	-0.00545	
24	Per: 5	-0.00512	

Idx	Property	mean SHAP	Distribution
25	Per: 6	-0.00364	
26	Per: 7	0.00000	
27	EN: [0.5, 0.85)	0.00000	
28	EN: [0.85, 1.2)	0.00000	
29	EN: [1.2, 1.55)	0.00299	
30	EN: [1.55, 1.9)	-0.00040	
31	EN: [1.9, 2.25)	-0.00799	
32	EN: [2.25, 2.6)	-0.00245	
33	EN: [2.6, 2.95)	0.00000	
34	EN: [2.95, 3.3)	0.00123	
35	EN: [3.3, 3.65)	0.00006	
36	EN: [3.65, 4]	0.00000	
37	CR: [25, 47.5) pm	0.00000	
38	CR: [47.5, 70) pm	0.00301	
39	CR: [70, 92.5) pm	-0.00011	
40	CR: [92.5, 115) pm	0.00080	
41	CR: [115, 138) pm	0.00061	
42	CR: [138, 160) pm	-0.01044	
43	CR: [160, 182) pm	0.00000	
44	CR: [182, 205) pm	0.00000	
45	CR: [205, 228) pm	0.00000	
46	CR: [228, 250] pm	0.00000	
47	VE: 1	0.00049	
48	VE: 2	0.00000	

(figure continued on next page)

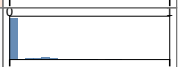
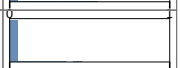
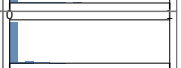
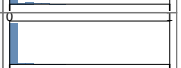
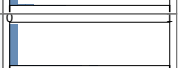
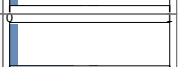
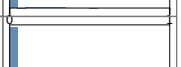
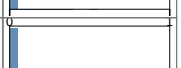
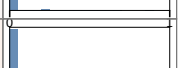
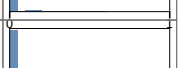
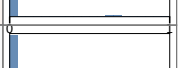
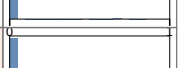
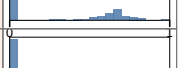
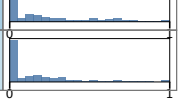
86

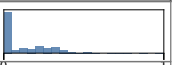
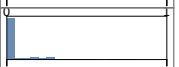
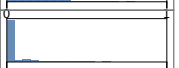
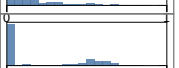
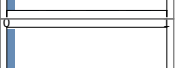
Idx	Property	mean SHAP	Distribution
49	VE: 3	0.00208	
50	VE: 4	-0.00104	
51	VE: 5	-0.00588	
52	VE: 6	-0.00354	
53	VE: 7	-0.00137	
54	VE: 8	-0.00128	
55	VE: 9	-0.00013	
56	VE: 10	0.00000	
57	VE: 11	0.00049	
58	VE: 12	0.00000	
59	FIE: [3.67, 4.48) eV	0.00000	
60	FIE: [4.48, 5.47) eV	-0.00000	
61	FIE: [5.47, 6.69) eV	0.00014	
62	FIE: [6.69, 8.17) eV	-0.00344	
63	FIE: [8.17, 9.97) eV	0.00022	
64	FIE: [9.97, 12.2) eV	-0.00119	
65	FIE: [12.2, 14.9) eV	0.00900	
66	FIE: [14.9, 18.2) eV	0.00000	
67	FIE: [18.2, 22.2) eV	0.00000	
68	FIE: [22.2, 27.1] eV	0.00000	
69	EA: [-3, -2.33) eV	0.00000	
70	EA: [-2.33, -1.66) eV	0.00000	
71	EA: [-1.66, -0.99) eV	0.00000	
72	EA: [-0.99, -0.32) eV	0.00000	

Idx	Property	mean SHAP	Distribution
73	EA: [-0.32, 0.35) eV	-0.00013	
74	EA: [0.35, 1.02) eV	-0.00636	
75	EA: [1.02, 1.69) eV	0.00482	
76	EA: [1.69, 2.36) eV	-0.00183	
77	EA: [2.36, 3.03) eV	0.00000	
78	EA: [3.03, 3.7] eV	0.00089	
79	Blk: s	0.00561	
80	Blk: p	0.00116	
81	Blk: d	-0.00334	
82	Blk: f	-0.00005	
83	AV: [4.48, 5.93) cm³/mol	-0.00010	
84	AV: [5.93, 7.85) cm³/mol	-0.00709	
85	AV: [7.85, 10.4) cm³/mol	-0.00011	
86	AV: [10.4, 13.7) cm³/mol	0.00997	
87	AV: [13.7, 18.2) cm³/mol	-0.00191	
88	AV: [18.2, 24) cm³/mol	-0.00536	
89	AV: [24, 31.8) cm³/mol	0.00000	
90	AV: [31.8, 42.1) cm³/mol	0.00000	
91	AV: [42.1, 55.7) cm³/mol	0.00000	
92	AV: [55.7, 73.7] cm³/mol	0.00000	

(figure continued on next page)

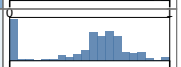
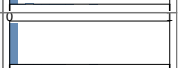
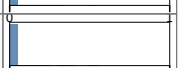
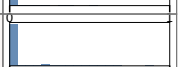
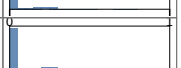
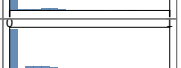
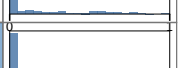
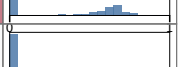
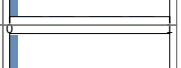
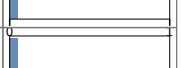
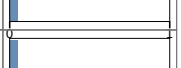
87

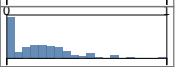
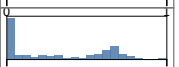
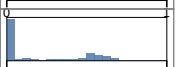
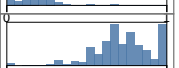
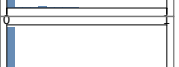
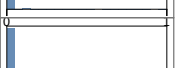
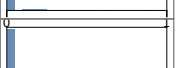
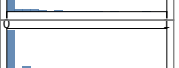
Idx	Property	mean SHAP	Distribution
93	Grp: 1	-0.00002	
94	Grp: 2	0.00256	
95	Grp: 3	0.00000	
96	Grp: 4	0.00015	
97	Grp: 5	-0.00018	
98	Grp: 6	-0.00030	
99	Grp: 7	0.00000	
100	Grp: 8	-0.00049	
101	Grp: 9	-0.00036	
102	Grp: 10	-0.00005	
103	Grp: 11	-0.00022	
104	Grp: 12	0.00043	
105	Grp: 13	0.00105	
106	Grp: 14	-0.00013	
107	Grp: 15	-0.00016	
108	Grp: 16	-0.00081	
109	Grp: 17	0.00002	
110	Grp: 18	0.00077	
111	Grp: unk	0.00000	
112	Per: 1	0.00000	
113	Per: 2	0.00204	
114	Per: 3	0.00151	
115	Per: 4	-0.00613	
116	Per: 5	-0.00434	

Idx	Property	mean SHAP	Distribution
117	Per: 6	-0.00595	
118	Per: 7	0.00000	
119	EN: [0.5, 0.85)	0.00000	
120	EN: [0.85, 1.2)	-0.00029	
121	EN: [1.2, 1.55)	0.00101	
122	EN: [1.55, 1.9)	-0.00011	
123	EN: [1.9, 2.25)	-0.01141	
124	EN: [2.25, 2.6)	-0.00073	
125	EN: [2.6, 2.95)	0.00000	
126	EN: [2.95, 3.3)	0.00142	
127	EN: [3.3, 3.65)	0.00023	
128	EN: [3.65, 4]	0.00000	
129	CR: [25, 47.5) pm	0.00000	
130	CR: [47.5, 70) pm	0.00259	
131	CR: [70, 92.5) pm	-0.00011	
132	CR: [92.5, 115) pm	0.00007	
133	CR: [115, 138) pm	-0.00021	
134	CR: [138, 160) pm	-0.01156	
135	CR: [160, 182) pm	-0.00007	
136	CR: [182, 205) pm	-0.00006	
137	CR: [205, 228) pm	0.00000	
138	CR: [228, 250] pm	0.00000	
139	VE: 1	0.00237	
140	VE: 2	0.00000	

(figure continued on next page)

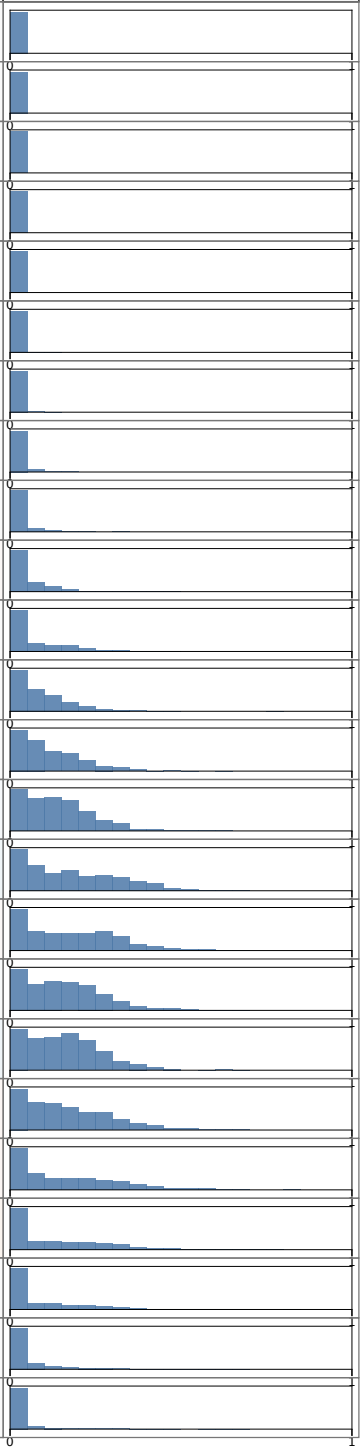
88

Idx	Property	mean SHAP	Distribution
141	VE: 3	0.00097	
142	VE: 4	-0.00126	
143	VE: 5	-0.00907	
144	VE: 6	-0.00184	
145	VE: 7	-0.00026	
146	VE: 8	-0.00034	
147	VE: 9	0.00000	
148	VE: 10	0.00000	
149	VE: 11	0.00022	
150	VE: 12	0.00000	
151	FIE: [3.67, 4.48) eV	0.00000	
152	FIE: [4.48, 5.47) eV	-0.00003	
153	FIE: [5.47, 6.69) eV	0.00000	
154	FIE: [6.69, 8.17) eV	-0.00462	
155	FIE: [8.17, 9.97) eV	0.00003	
156	FIE: [9.97, 12.2) eV	-0.00882	
157	FIE: [12.2, 14.9) eV	0.01007	
158	FIE: [14.9, 18.2) eV	0.00000	
159	FIE: [18.2, 22.2) eV	0.00000	
160	FIE: [22.2, 27.1] eV	0.00000	
161	EA: [-3, -2.33) eV	0.00000	
162	EA: [-2.33, -1.66) eV	0.00000	
163	EA: [-1.66, -0.99) eV	0.00000	
164	EA: [-0.99, -0.32) eV	0.00000	

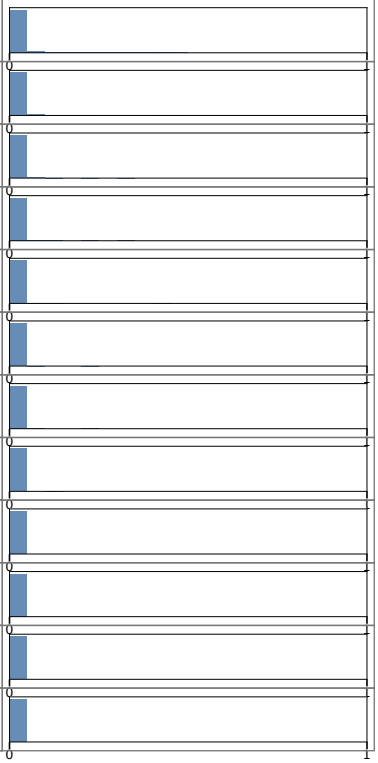
Idx	Property	mean SHAP	Distribution
165	EA: [-0.32, 0.35) eV	-0.00056	
166	EA: [0.35, 1.02) eV	-0.00554	
167	EA: [1.02, 1.69) eV	0.00499	
168	EA: [1.69, 2.36) eV	-0.00231	
169	EA: [2.36, 3.03) eV	0.00000	
170	EA: [3.03, 3.7] eV	0.00087	
171	Blk: s	0.00276	
172	Blk: p	-0.00021	
173	Blk: d	-0.00328	
174	Blk: f	-0.00002	
175	AV: [4.48, 5.93) cm³/mol	-0.00017	
176	AV: [5.93, 7.85) cm³/mol	-0.00180	
177	AV: [7.85, 10.4) cm³/mol	-0.00002	
178	AV: [10.4, 13.7) cm³/mol	0.00477	
179	AV: [13.7, 18.2) cm³/mol	-0.00093	
180	AV: [18.2, 24) cm³/mol	-0.00399	
181	AV: [24, 31.8) cm³/mol	0.00000	
182	AV: [31.8, 42.1) cm³/mol	0.00000	
183	AV: [42.1, 55.7) cm³/mol	0.00000	
184	AV: [55.7, 73.7] cm³/mol	0.00002	

(figure continued on next page)

89

Idx	Property	mean SHAP	Distribution
185	0.00-0.20 Å	0.00000	
186	0.20-0.40 Å	0.00000	
187	0.40-0.60 Å	0.00000	
188	0.60-0.80 Å	0.00000	
189	0.80-1.00 Å	0.00000	
190	1.00-1.20 Å	-0.00001	
191	1.20-1.40 Å	0.00000	
192	1.40-1.60 Å	0.00000	
193	1.60-1.80 Å	0.00000	
194	1.80-2.00 Å	0.00024	
195	2.00-2.20 Å	0.00000	
196	2.20-2.40 Å	-0.00003	
197	2.40-2.60 Å	0.00000	
198	2.60-2.80 Å	-0.00040	
199	2.80-3.00 Å	-0.00102	
200	3.00-3.20 Å	0.00019	
201	3.20-3.40 Å	-0.00017	
202	3.40-3.60 Å	-0.00109	
203	3.60-3.80 Å	0.00060	
204	3.80-4.00 Å	0.00056	
205	4.00-4.20 Å	0.00081	
206	4.20-4.40 Å	0.00003	
207	4.40-4.60 Å	0.00000	
208	4.60-4.80 Å	0.00000	

(figure continued on next page)

Idx	Property	mean SHAP	Distribution
209	4.80-5.00 Å	0.00000	
210	5.00-5.20 Å	0.00000	
211	5.20-5.40 Å	0.00000	
212	5.40-5.60 Å	0.00000	
213	5.60-5.80 Å	0.00000	
214	5.80-6.00 Å	0.00000	
215	6.00-6.20 Å	0.00000	
216	6.20-6.40 Å	0.00000	
217	6.40-6.60 Å	0.00000	
218	6.60-6.80 Å	0.00000	
219	6.80-7.00 Å	0.00000	
220	7.00-7.20 Å	0.00000	

Supplementary Figure 8. Average SHAP value and distribution for SVR across all segments. The feature abbreviations are defined as follows: Grp (group number), Per (period number), EN (electronegativity), CR (covalent radius), VE (valence electrons), FIE (first ionization energy), EA (electron affinity), Blk (block), and AV (atomic volume). Features are grouped into three segments: self-atom features (Idx 1–92), neighbor-atom features (Idx 93–184), and encoded distance features (Idx 185–220). In each row, the Distribution panel shows the distribution of training-set values for the corresponding feature on the normalized 0–1 input scale; taller bars indicate that more training samples fall within that interval.