Supplemental Information for The Illusion of Chemical Continuity Limits Machine Learning Generalization

(Dated: 23 July 2026)

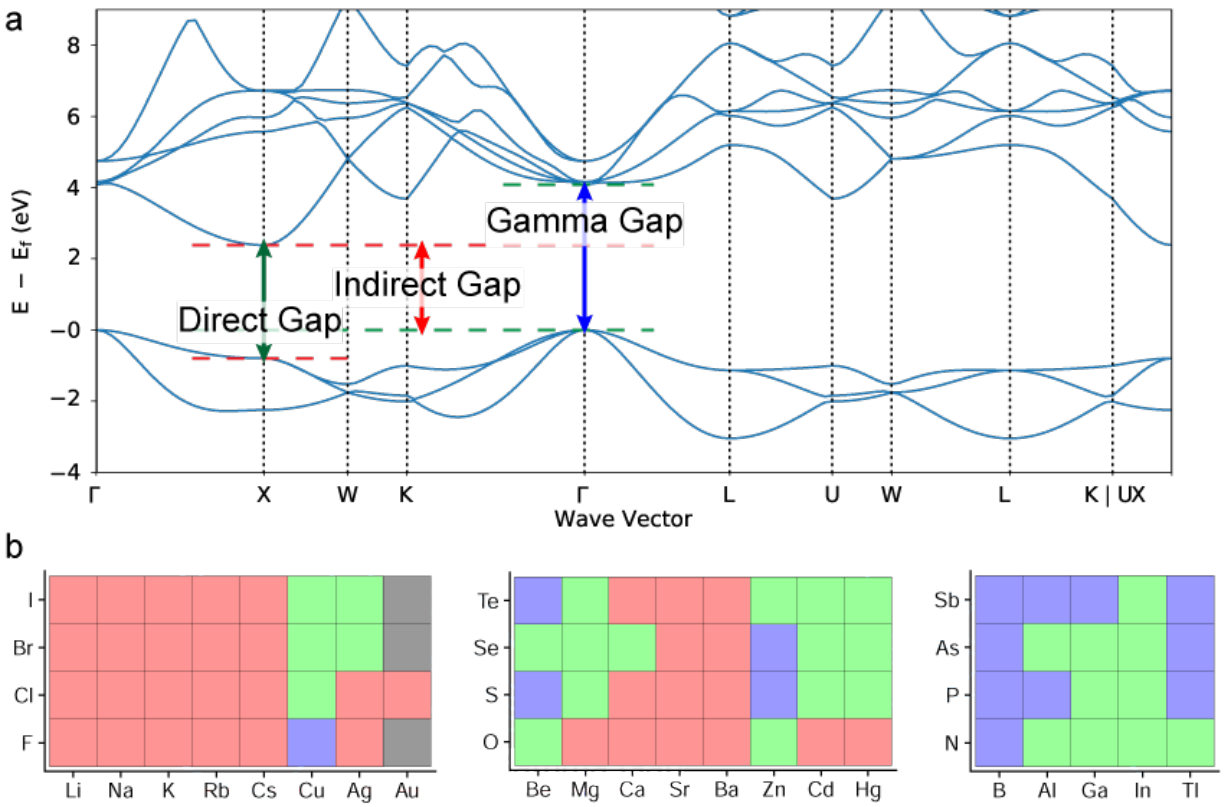

FIG. S1. **a,** Bandstructure of CaS (mp-1672) highlighting the differences in indirect, direct, and gamma gaps.[1] Indirect and direct gaps exist in different k-points in contrast to gamma gaps. Panel **b** shows the zincblende and rocksalt structure coverage map in Materials Project with blue as zincblende, red as rocksalt, green as both, and grey as none.

I. SUPPLEMENTAL INFORMATION

A. Additional Structures

84 zincblendes were computed to obtain gamma gaps, of which 40 were not in the materials project database (Fig. S1). This was done in hopes of covering the unrepresented 1-7 and 2-6 cations and maintaining consistency in computational methods between the materials considered. The gamma gap subset of Materials Project contained 38 structures that were identified as zincblendes (cubic with four neighbouring atoms at each site).

An additional 84 CsCl structures were also computed and compared. The comparisons between CsCl, rocksalts, and zincblendes 1-7 and 2-6 materials are outlined in figure S2. The comparisons are similar to those between rocksalts and zincblendes with no obvious

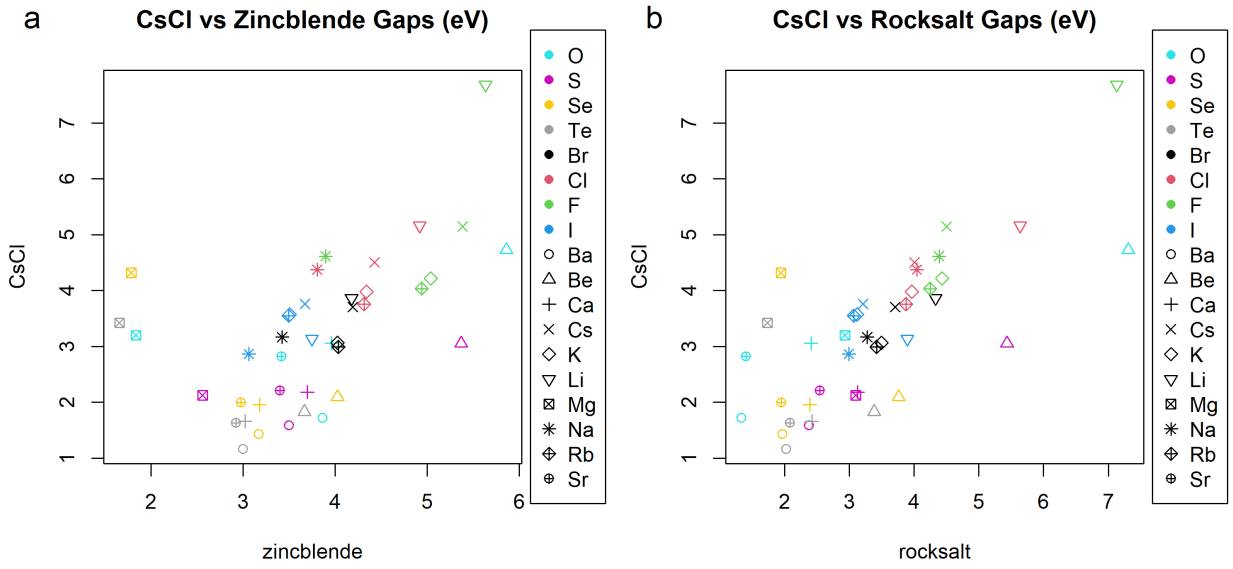

FIG. S2. A comparison of CsCl against zincblende (a) and rocksalt(b) gamma gaps. The scatters exhibit no apparent trend with R^2s of 0.3 and 0.5 respectively.

trends beyond general composition (Fig. 1).

The 3D representations of the gamma gaps are also substantially different from those of the rocksalt and zincblende materials (Fig. S3). Key oddities are Mg and Na as before.

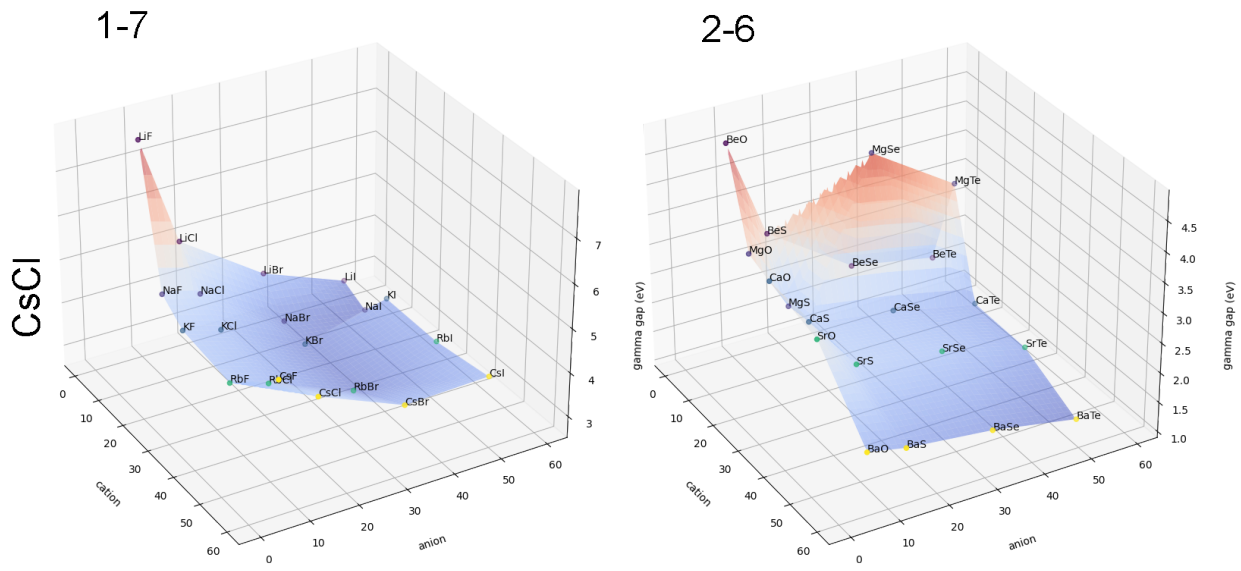

FIG. S3. Visualization of the gamma gaps of the 1-7 and 2-6 CsCl materials across cation and anion identity. Pairing the compounds by atom row (proxied by atomic number) displays notable deviations in slope by both axes similar to that of rocksalts and zincblendes.

B. Reducing Model Complexity

As newer models do not seem to drastically improve zincblende prediction, it was postulated that the structural representational components of the GNN may be to blame for poor performance. CoGN acting as the best bandgap predictor at this time for materials project bandgaps does not appear to perform differently from CGCNN-styled models (Fig. S4).[2,3] This led to the attempts to remove any confounding issues that the encoding process may introduce to the model.

A GNN can be roughly broken down into the graph convolution modules, the pooling module, and the main feed-forward layers (Fig. S5). Graph convolution generates a position-independent representation of the material of indefinite length. Pooling graph features fixes the vector length by applying a function to collate all atoms into a unified vector. Typically, this involves averaging all convolved atoms in the graph which generates a uniform "representative" atom and its associated bonds.

However, this pooling step introduces unavoidable information loss for which magnitude is uncertain. Reducing the available crystal structures to a single type bypasses this need by fixing the number of atoms and bonds. By considering only binary compounds, the

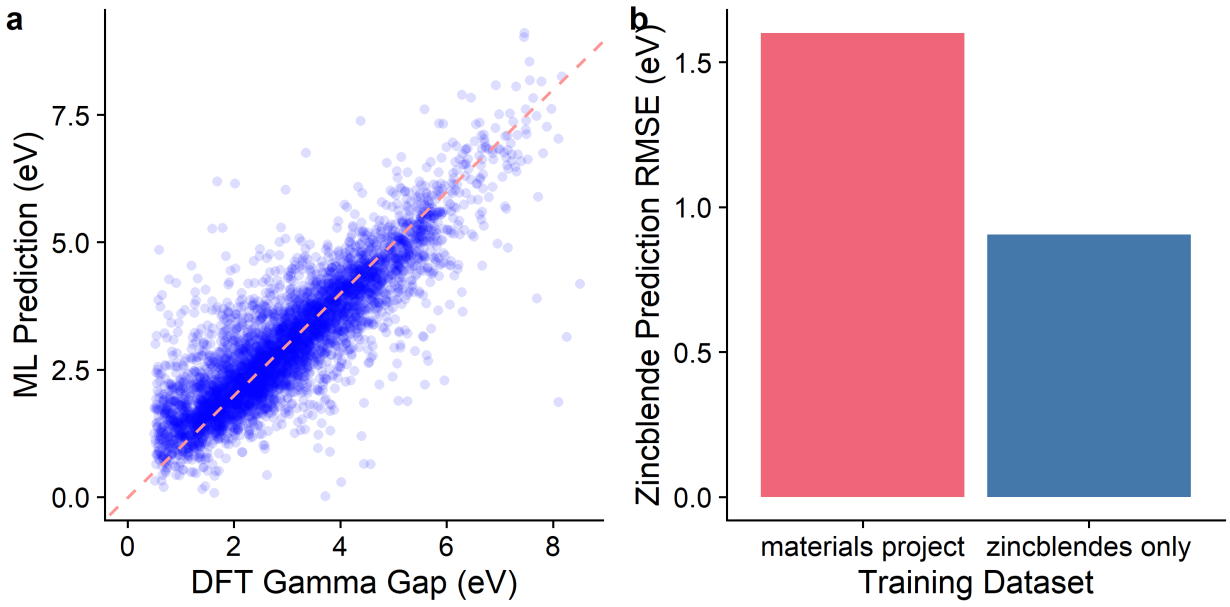

FIG. S4. **a,** CGCNN performance trained using gamma gaps is similar to the bandgap predictions reported in literature.[2] **b,** Predictions on a testing split of zincblendes (red) are substantially higher than that of training solely on zincblendes (blue), despite the introduction of other structures from the database.

order of the atoms in the listing can be standardized to cation and anions. So long as the atomic order is maintained, there is no additional complexity beyond atom identity and bond properties needed. Zincblendes are one such case with symmetrical coordination between the two atomic species. Considering their underrepresentation in Materials Project, they made for a convenient test case for assessing model performance (Fig. S1b).

To maintain a sensible cation-anion ordering, the materials under consideration must contain a distinct cation and anion. Limiting the atoms to s- and p-block elements avoids the complexities associated with transition metals, such as multiple oxidation states keeping things consistent. These restrictions lead to combinations of group 1, 2, and 13 cations combined with group 15, 16, and 17 anions. The combinations will herein be categorized by their valence electron counts: 1-7 for halides, 2-6 for chalcogenides, and 3-5 for pnictogenides.

To avoid the complexities associated with radioactive compounds, the atomic number was capped at 82 for combinations under consideration. Hydrogen was also avoided due to its propensity to be both a cation and an anion in conjunction with the group 1 and 17 atoms. This leaves five cations from each group and four anions giving a total of 3 sets of 20

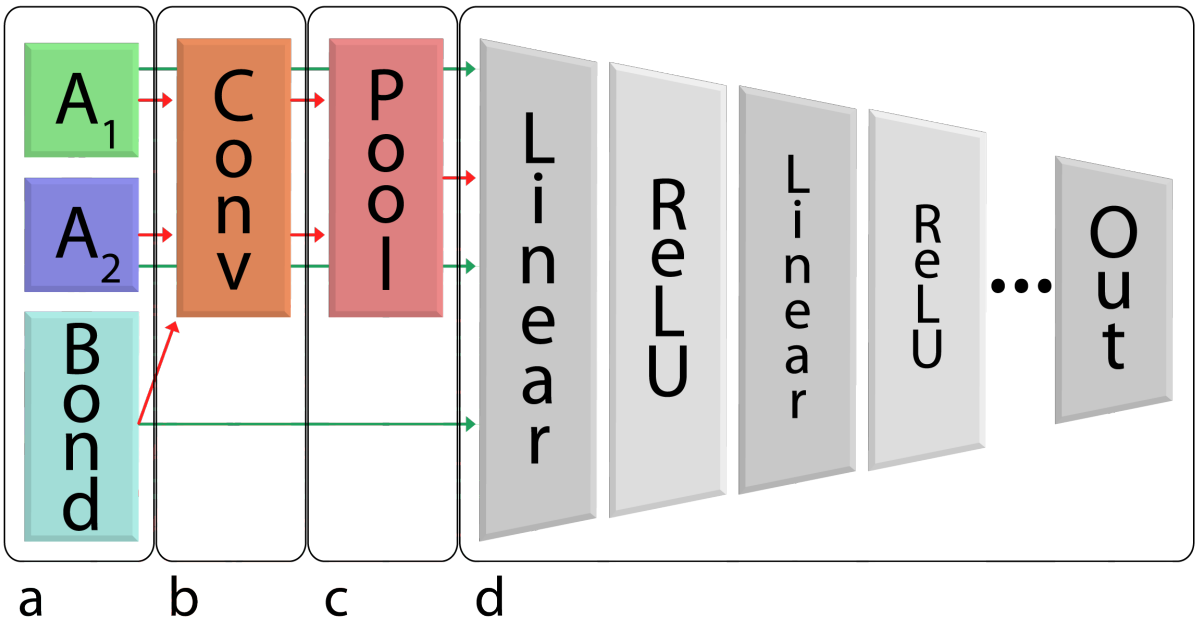

FIG. S5. An illustrated diagram of the model architecture and the component pieces: **a,** model input consisting of atoms and bond features, **b,** graph convolution, **c,** pooling, and **d,** the main feed forward linear layers. Removal of distinct crystal structures allows for the removal of components b and c provided the input stream now accommodates any bond properties expected from graph convolution (from red arrows to green arrows). The resulting, heavily simplified, model is therefore comprised of a and d only.

compounds, which drastically reduces the sample size. Any attempts to increase this sample size is now limited by the size of the periodic table itself.

As additional crystal structures do not appear to meaningfully improve training, the model overhead associated with encoding different crystal structures is similarly removable. This allows for the contributions to accuracy afforded by model architecture to be assessed (Fig. S6).

Stepwise component removal leads to a further five distinct models depending on the removed components (Fig. S5): the pooled graph with all features intact, the static graph where the pooling layer is removed, the pooled model where the graph component (connectivity) is removed, the convolved model where the pooling and graph are removed, and the forwards feeding model where only the essential neural network remains.

The validation RMSEs for wurtzites are shown in Figure S6a. Remarkably, prediction

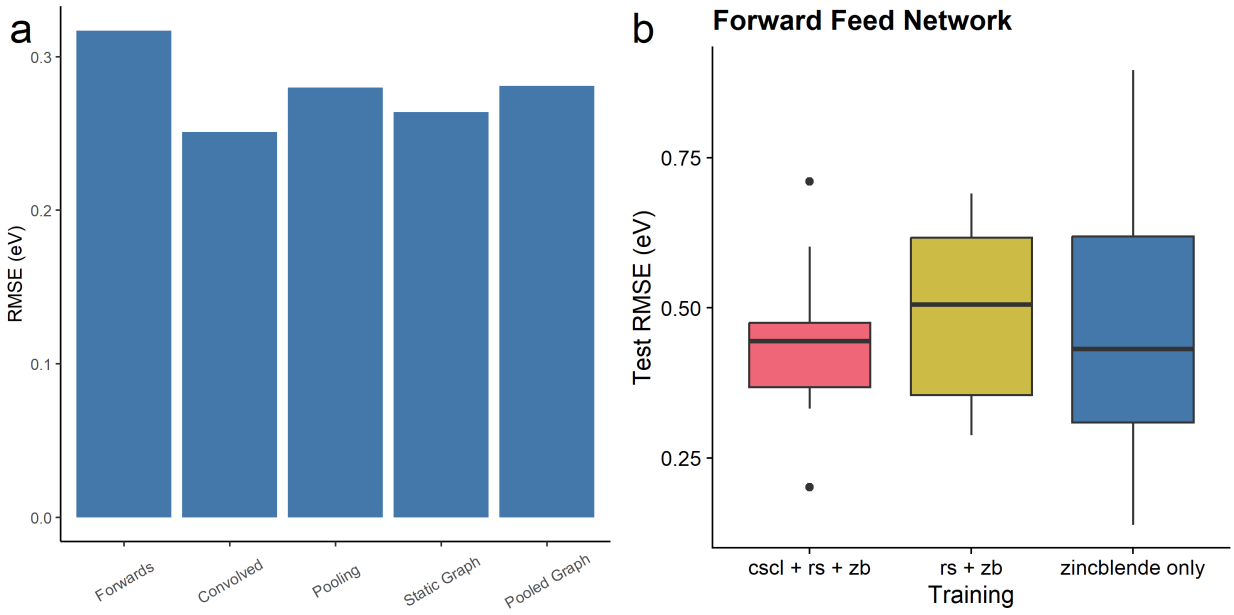

FIG. S6. **Architectural ablation study. a,** Validation comparison between various model architectures based off a single training-validation split. The bandgap losses are generally unchanged between model complexity shifts with major shifts where pooling steps are introduced. **b,** The performance on test zincblendes of the minimal forward feeding model. Even with this minimal architecture, adding structural diversity (CsCl + rocksalts) increases prediction error compared to training on zincblendes alone.

accuracy did not degrade significantly with reduced complexity, suggesting that structural information is generally ignored by the GNNs.

Considering the minor gains from adding graphs, addressing the underlying RMSE was deemed more important. The lowered losses in the models stem mostly from careful splitting and optimization. Wurtzites were originally used here before migrating to zincblendes due to their improved representation in the Materials Project database.[4]

Adding alternative crystal structures to zincblende, even within the fully reduced feed-forward network, does not substantially improve predictive performance (Fig. S6b). The model was trained using the same fixed zincblende split as in the main text, with rocksalt and CsCl structures added to the training set as appropriate (84 combinations per split). Although marginal gains are observed with the inclusion of CsCl structures, prediction errors for the MgO and NaF test cases remain largely unchanged. Instead, it appears that the additional structures seem to clamp the extremities towards some shared average.

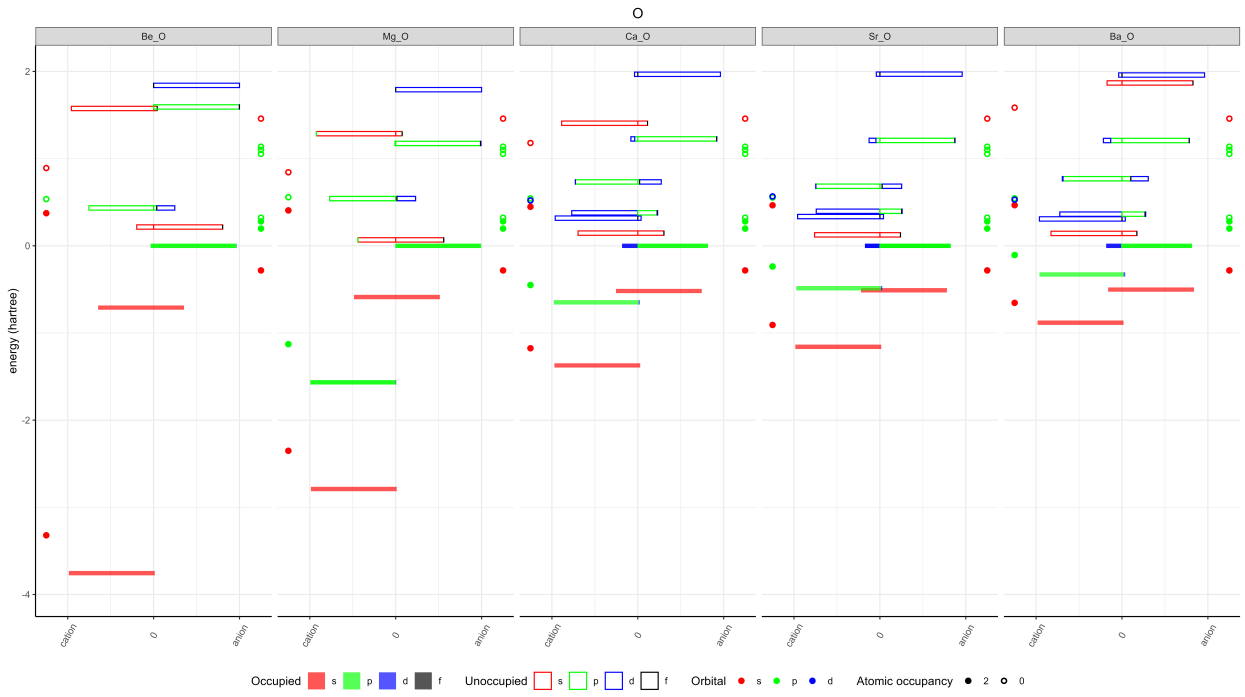

FIG. S7. Projected PDOS of zincblende oxides. The bandgap is situated near the zero line between the top occupied and bottom unoccupied orbitals. The orbital contributions are by colour with red for s, green for p, blue for d, and black for f. The points on the side represent the component atomic orbitals relative to each other.

C. Bandgap components

Visualizing the PDOS of the crystals (CP2K) shows some of the hidden complexities in bandgap which the periodic table introduces (Fig. S7). The major shift comes from the introduction of d-orbitals at period 4 elements. This seems to increase the bandgap splitting from the occupied p-p split and an unoccupied s-s split for oxides. This pattern is seen in most of the other combinations as well. Of note, most of the molecular orbitals appear to be purely of said type with little inter-l mixing for the s and p orbitals.

This suggests that the periodic table is not just fragmented by group, but also by the introduction of d orbitals. Of course, this functionally reduces the number of samples for any given trend.

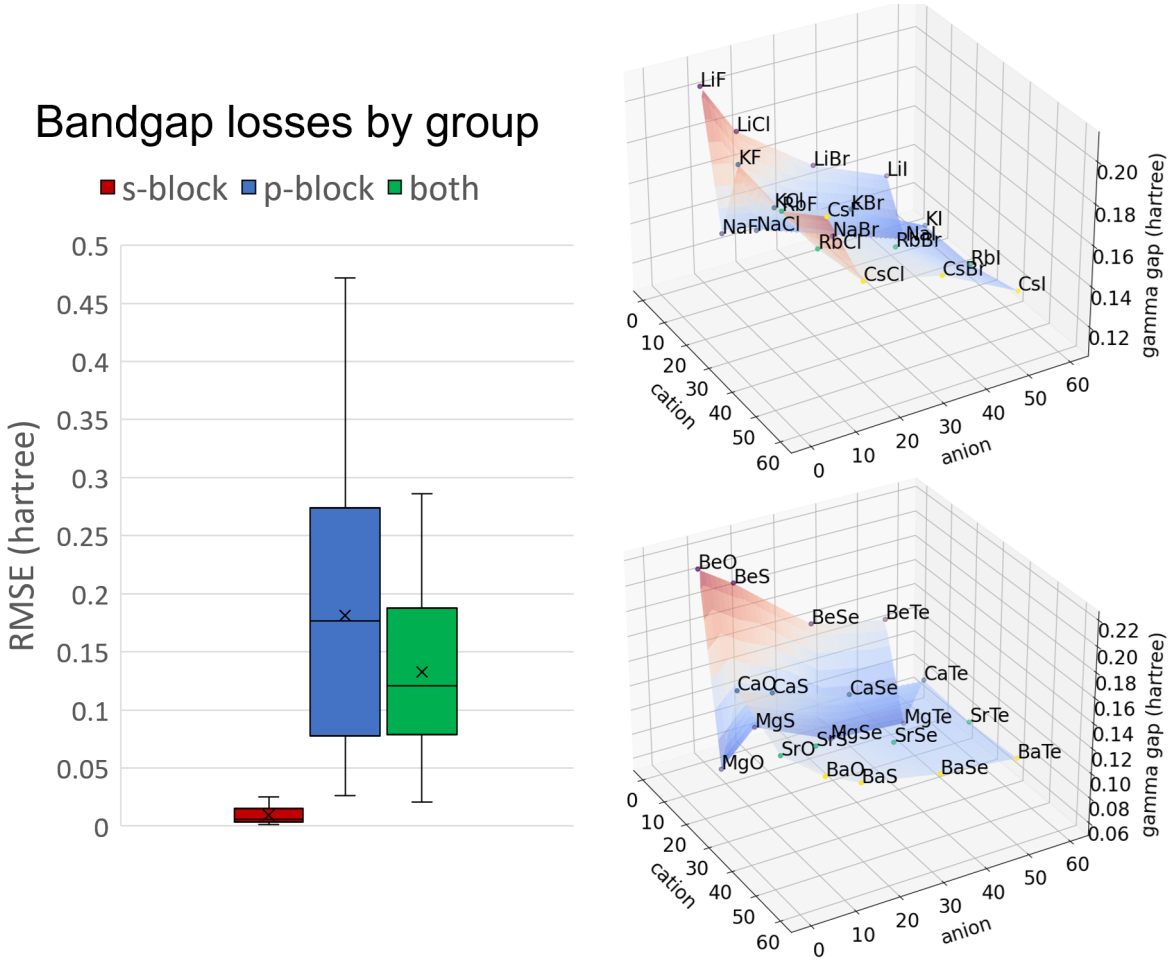

FIG. S8. The major shift only occurs when isolating the groups from each other, showing that for the restricted dataset, **a,** the majority of the difficulty lies in the pnictogenide bandgaps which are substantially different from the s-block cations. **b,** Visualizing the gamma gaps of the 1-7 and **c,** 2-6 zincblende materials across cation and anion identity show significant differences between the surfaces.

D. Groupwise Dependency

Initial modified model trials, however, were disappointing. Removing crystal structure complexity and the associated graph convolution did not substantially reduce predictive loss (Fig. S8). Varying the combinations for atomic input features as well as explicit atomic orbital energies from pseudopotential input files also failed to significantly improve performance.[5–8]

A significant shift was observed by removing the 3-5 materials from the training set but

the remaining 1-7 and 2-6 materials still predicted poorly. However, with the dataset at this size, it was now humanly possible to investigate the samples individually to find any explanation to the noise. Plotting the gamma gaps against the various properties of their component atoms revealed a disturbing trend: that there was no real trend at all (Fig. S8).

E. Parameter Combinations

Hyperparameters were initially tuned via Weights and Biases minimizing validation losses.[9] The parameters of concern were as follows: learning rate, weight decay, layer width, network depth, layer size changes, and learning rate schedules. However, due to the small dataset size, the hyperparameters did not always lead to a stable minimum based on split. This led to the manual refinement of learning rates and weight decays from automated minima.

The discontinuities of the periodic table are technically explainable since they are based on quantum mechanics. Notably, the filling of subshells appears to be the major delineation as in the singly occupied to doubly occupied p orbitals for the transition between N and O. Taking these into account for ML did not substantially affect the loss though through orbital definitions (Fig. S9 valence containing models). This is attributed to the increasingly smaller sample size presented by the fragmentation by group and by row that prevents ML from encoding a trend therein.

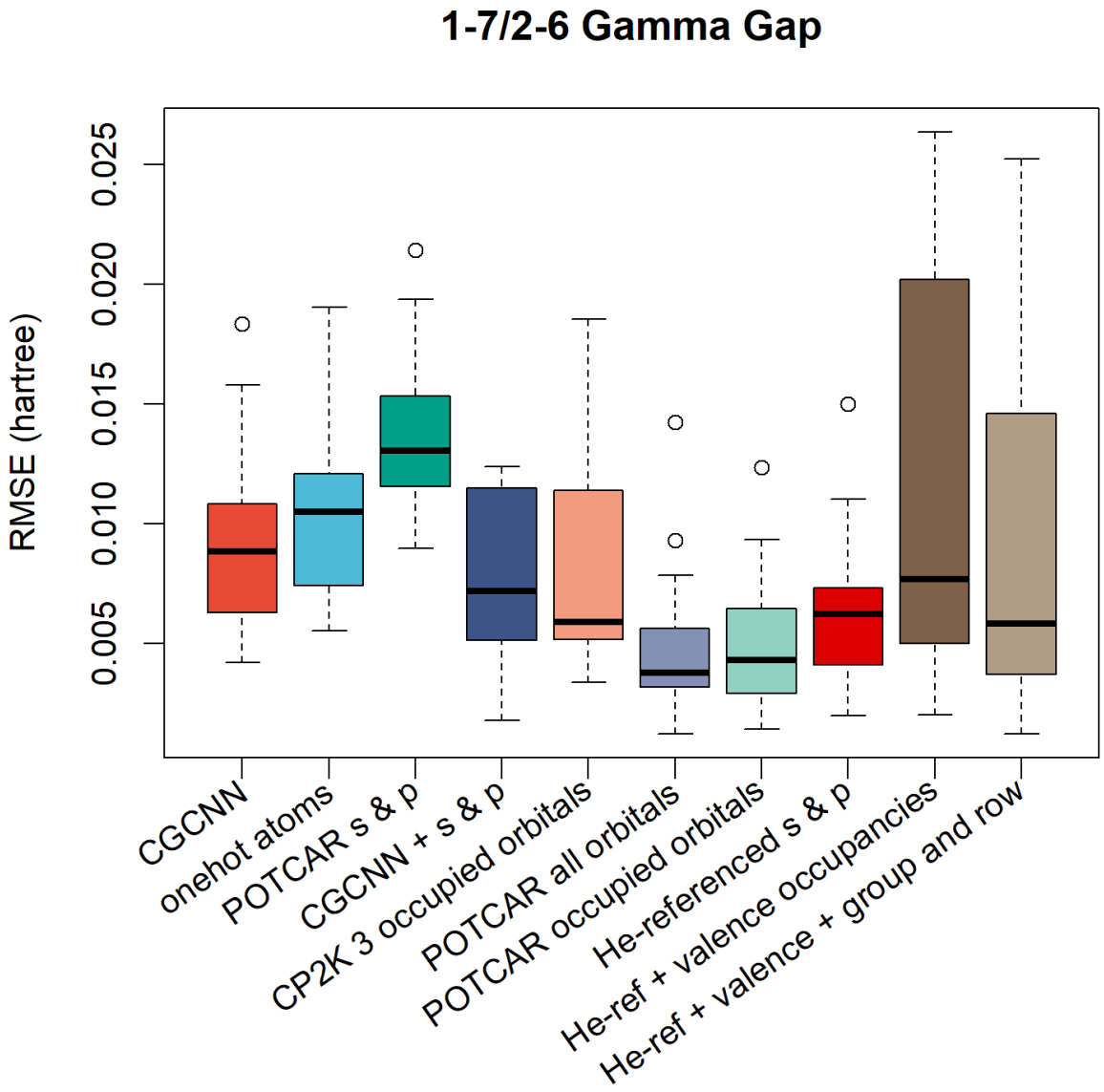

FIG. S9. Parameter combinations comparison for zincblende 1-7 and 2-6 materials. All models were trained/validated on 20 identical splits with different atomic parameters.

F. Periodic table trendlessness

Electronegativities also exhibit the discontinuities seen in ionization energies (Fig. S10). The breaks in trend appear to occur at similar locations. Atomic radii are smoother and thus not visualized.

Attempts to find trends in ionization energies are seen in figure S11. A linear model

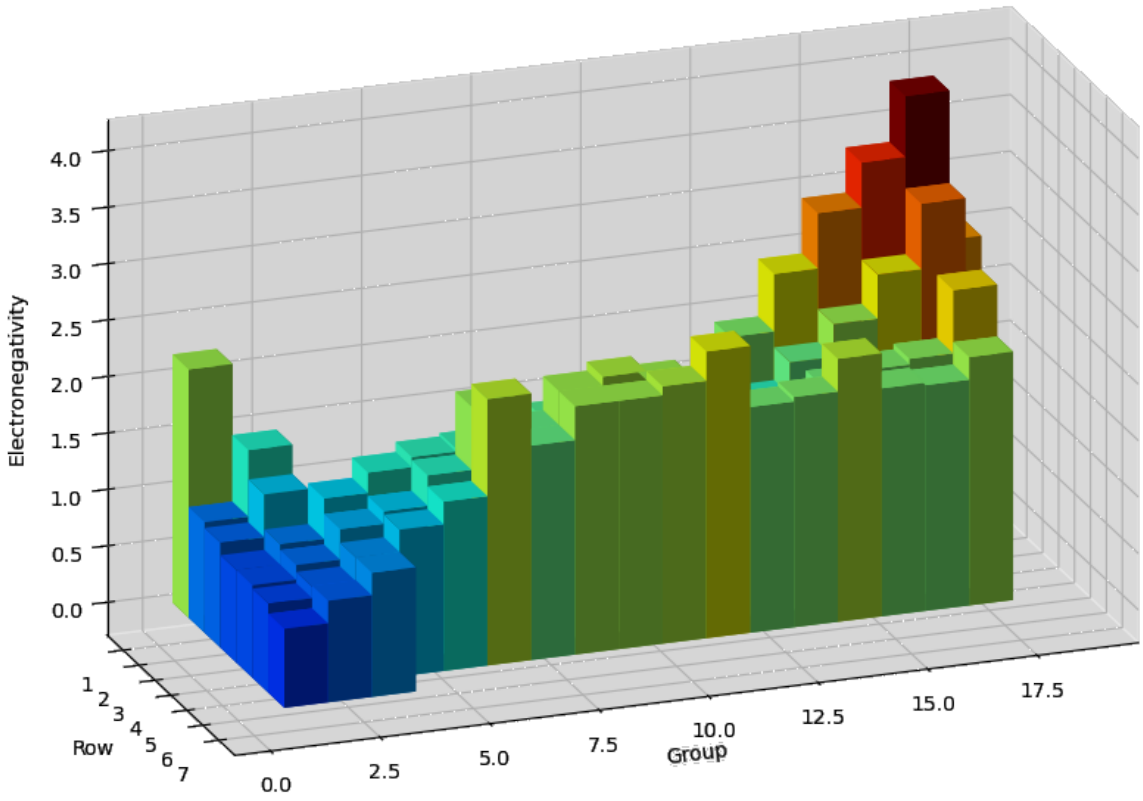

FIG. S10. The electronegativities of the periodic table represented as a bar plot sans f-block elements. Despite expectations, the periodic table's trends are not uniform as a whole. Similar to the ionization energies, discontinuities are seen between groups and rows.

was fit to the s and p block elements as a function of valence occupancy and row (group and period). The residuals were presented to search for any emergent trends in the non-linear component of the model. Considering that while the trends seem to flip around the period 4 elements, there is no apparent relation for the skew of the traces per row, making extrapolation from one row to another difficult.

ML models taking the filling of partial subshells into consideration performed similarly to the others. This suggests that the row-based complexity contributes to preventing clean mapping of valence to gaps.

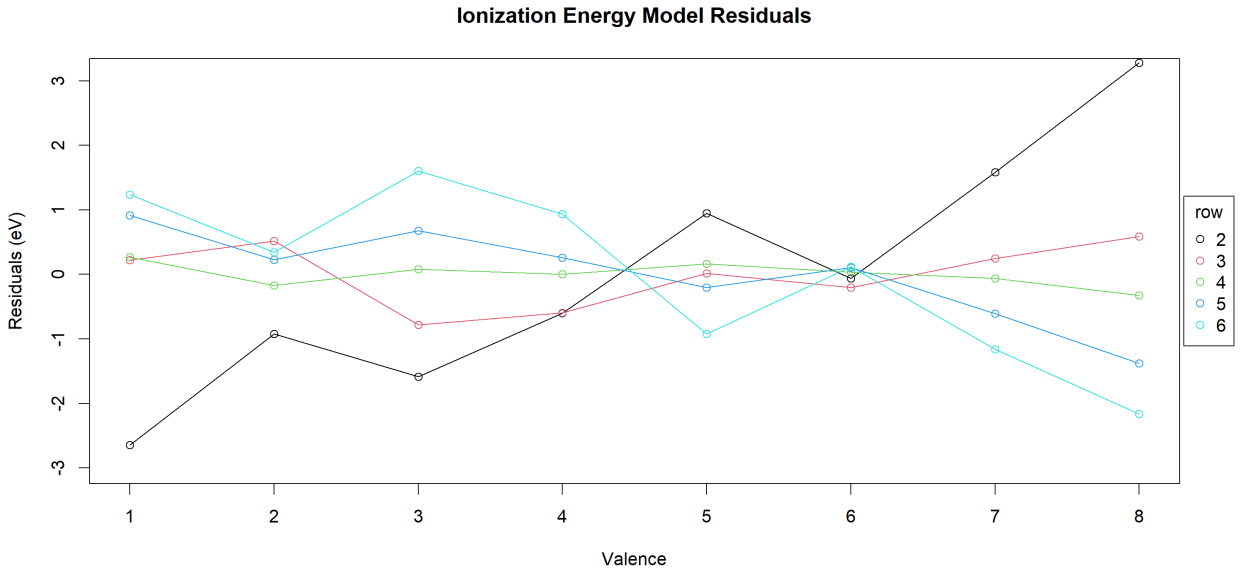

FIG. S11. Residuals of a linear model fit to the s and p block ionization energies as a function of valence occupancy and row.

G. Rocksalts versus zincblendes

A closer look at the residuals of the composition linear model for rocksalts versus zincblendes shows that there are multiple regimes in the clustering. These clusterings

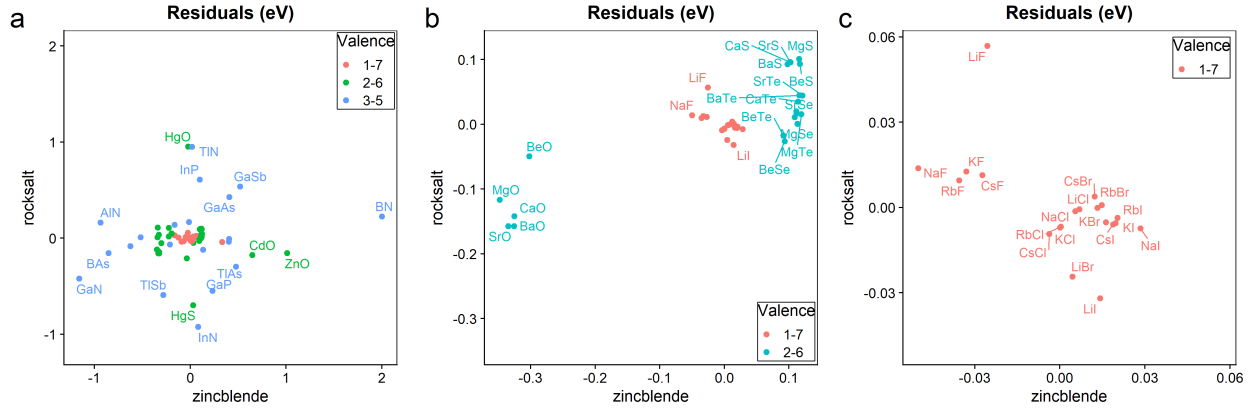

FIG. S12. A comparison of rocksalt and zincblende gamma gap residuals removing compositional influence. **a,** Coloured by combination, the materials appear to organize into shells. **b,** Zooming into the origin, the 2-6 materials show a substantial split due to the oxides. **c,** A similar split is observed for the 1-7 materials upon closer inspection for fluorides.

do not exhibit a greater general trend but instead reaffirm the difficulties for generalization beyond the stated crystal structures.

The general organization appears to correspond to multiple rings of residuals with 1-7s at the center and 3-5s at the edges of the domain. This appears to correspond with anticipated ionicities of the materials, and by extension, the expected splitting degree.

The 2-6 materials are especially odd, with oxides all in their own cluster away from the others. Otherwise, the relative positions of the materials appear random given the few datapoints available. As for the 1-7's a similar clustering of fluorides is seen with a disorganized remainder. The discrete scatters and sectioning by atom identity further reinforce the discreteness of the trends available, which complicates training.

H. Local roughness and predictability

This discontinuity is further highlighted when analyzing the distribution of prediction errors across the full Matbench dataset (Fig. S13b-f).[2] For bandgaps, the Principal Component Analysis (PCA) projection of the latent space (Fig. S13c) and its corresponding prediction error (Fig. S13d) reveal a "salt-and-pepper" distribution. High-error samples frequently appear as immediate nearest neighbours to low-error samples. This lack of spatial clustering in the error signal indicates that while the model successfully groups chemically similar materials, the bandgap property itself exhibits intrinsic, high-frequency discontinuities within these clusters.

In contrast, the formation energy manifold (Fig. S13e) exhibits smoother colour gradients. Crucially, the formation energy errors (Fig. S13f) are localized, with high-error regions clustered together. This shows that formation energy forms a smooth, learnable landscape where neighbour properties are predictive, whereas the bandgap landscape is dominated by local discontinuities that disrupt the model's ability to generalize, even between nearest neighbours in the learned representation.

This is reinforced by comparing the relative roughness between the target and the 5 nearest neighbours by embedding (Fig. S14). While both bandgaps and formation energies have a distinct, low noise axis where the distance to the local average does not drastically affect the error, a secondary trend can be seen in bandgaps that deviates from this principal axis which is not present in formation energies (a). This shows that a significant subset of

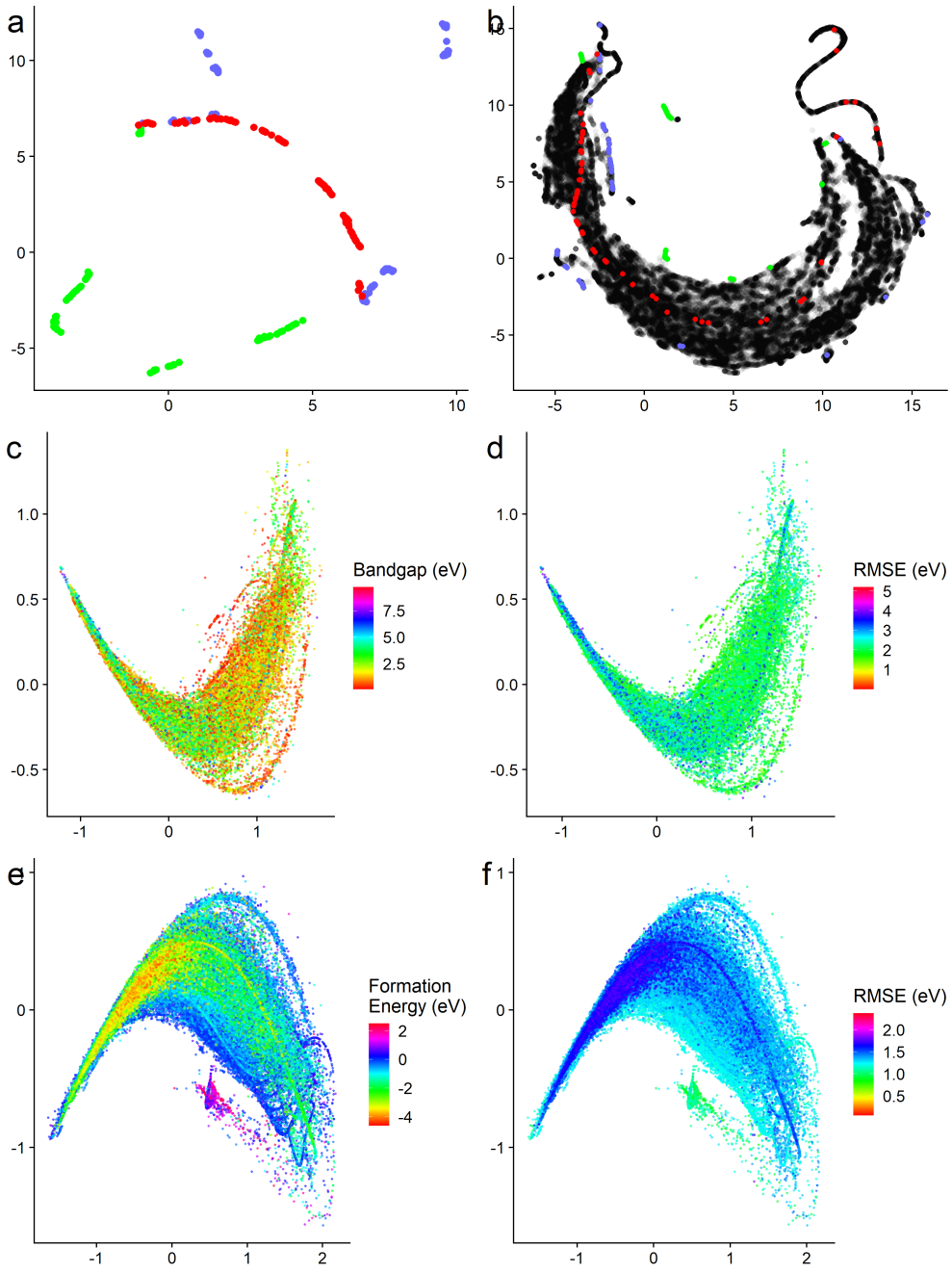

FIG. S13. Latent space visualization of the coGN model embeddings. **a,** UMAP projection of the binary test sets shows that rocksalt (red), zincblende (green), and CsCl (blue) structures occupy disjoint manifolds with no overlap.[10] **b,** Projection of the wider Materials Project dataset shows a complex, fragmented chemical space. **c-f,** Comparison of the latent space coloured by property value and prediction error. **d,** The bandgap error distribution exhibits a "salt-and-pepper" texture, where high-error points are intermixed with low-error neighbours, indicating a non-smooth landscape. **f,** In contrast, formation energy errors are more spatially clustered, indicating a smooth, learnable manifold.

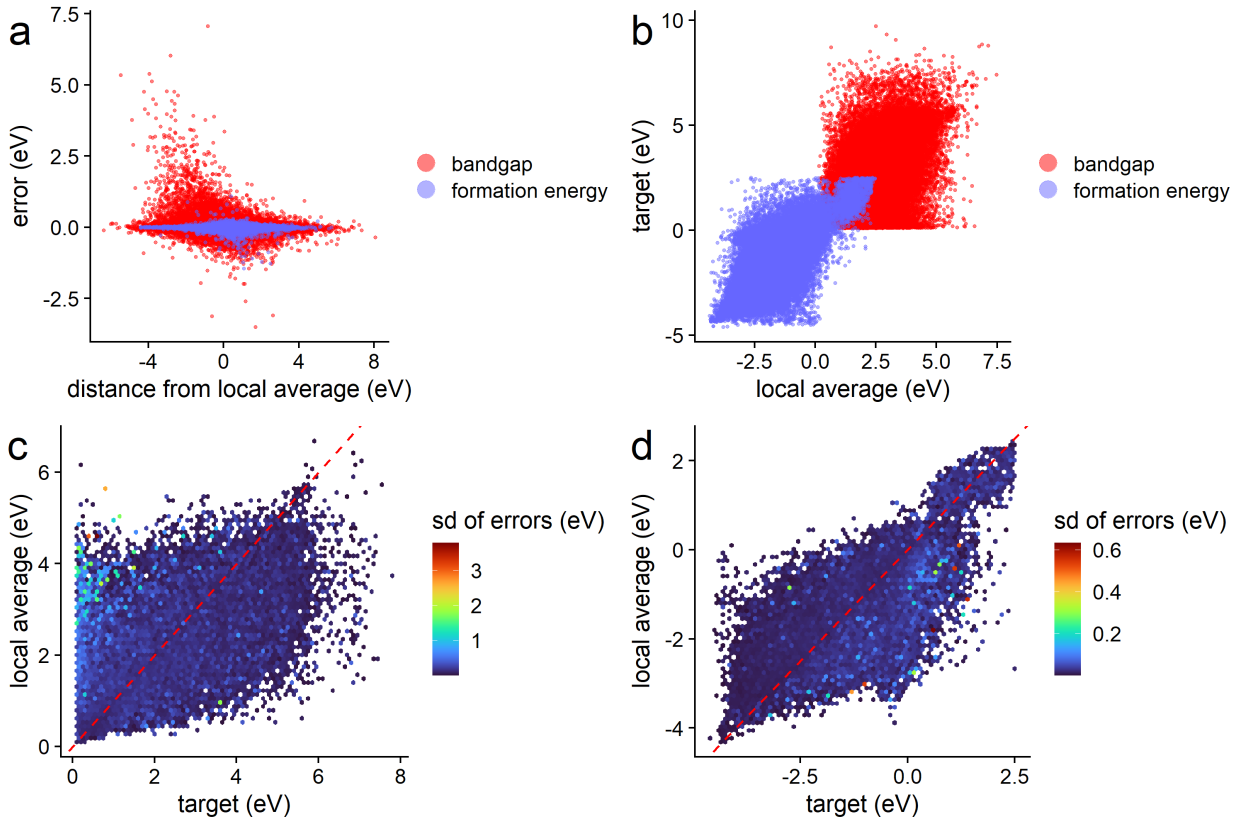

FIG. S14. **a,** Comparison between errors and the distance from the average of the 5 nearest neighbours for bandgaps (red) and formation energies (blue) shows bandgaps having a significantly larger error and roughness. **b,** This spread is also seen in how the target aligns with the local average. **c,** Looking at the standard deviation of errors for bandgaps versus **d,** formation energies shows that consistency falls at rougher parts of the manifold.

bandgaps is particularly sensitive to roughness which is not present in formation energies.

A similar comparison is visible in Magpie embeddings to a lesser extent suggesting that the issue is not isolated to ML alone(Fig. S15).[11,12] Instead, the disjoints are fundamental to the underlying manifold by multiple projections.

Mapping the errors to the relative roughness deviations shows that both formation energies and bandgaps exhibit higher error spreads at the extremities of the roughness eccentricity but as formation energies are spread less, their errors also end up being less severe.

When projected against atomic energies, the total energies form a mostly smooth plane (Fig. S16a). This is transferable between the 1-7 and 2-6 groups which lie on the same surface. In effect, total energies are directly correlated with atomic energies leading to

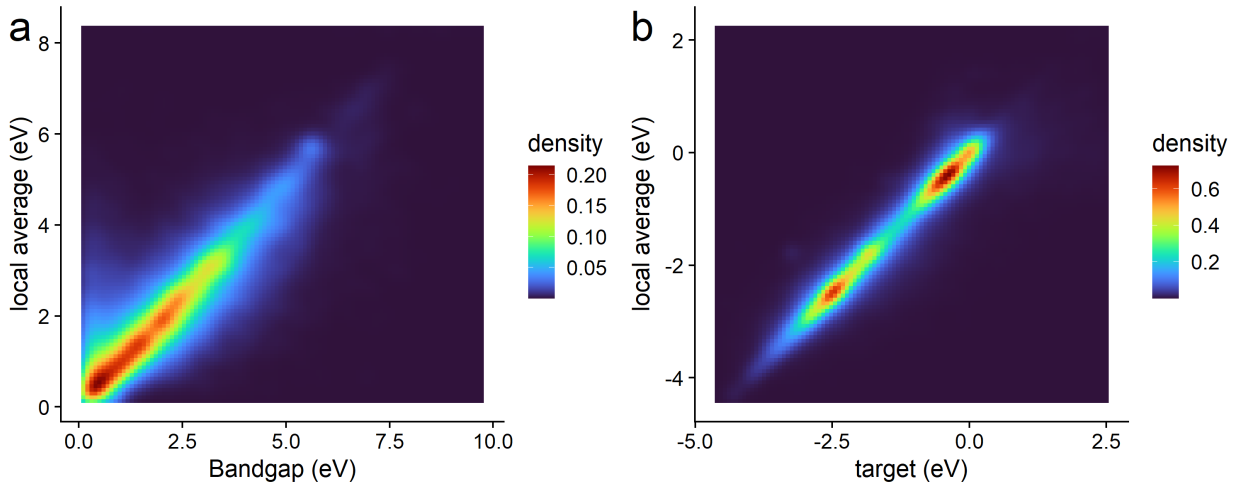

FIG. S15. **a,** A visualization of the predictive prowess of the 5 nearest neighbours via Magpie to the actual bandgap displays a notable spike near zeros that is **b,** absent in the better performing formation energies.

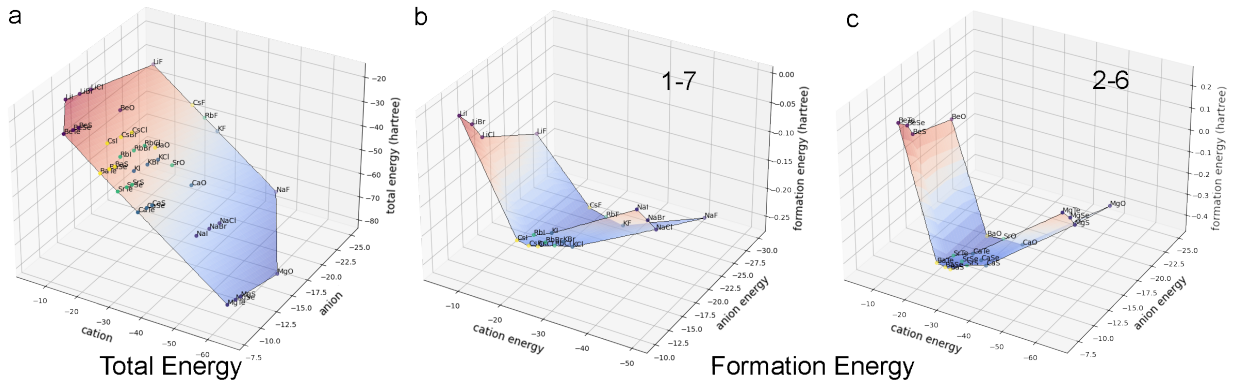

FIG. S16. 3D projections of total and formation energies of zincblendes. **a,** Total energies are functionally linear with respect to atomic energies. **b,** Formation energies for 1-7 and **c,** 2-6 zincblendes are nonlinear but monotonic with a dip at the heavier elements. Notably the atomic energies are not in order of the periods in this projection.

easier modelling.

The deviations to the plane are therefore the formation energies which are not as linear (Fig. S16b, c). The values projected against atomic energies are paraboloid but notably the lines mapped by the cation and anions are self-similar and can easily be mapped to each other. The surface is relatively smooth and monotonic.

Sets	D	p
Rocksalt vs zincblende gamma gaps	0.275	0.09707
Rocksalt vs zincblende residuals	0.3	0.05414
Zincblende 1-7 vs 2-6 gamma gaps	0.55	0.003967
Zincblende 1-7 vs 2-6 residuals	0.75	9.547×10^{-6}
Rocksalt 1-7 vs 2-6 gamma gaps	0.7	5.569×10^{-5}
Rocksalt 1-7 vs 2-6 residuals	0.5	0.0123

TABLE S1. Kolmogorov-Smirnov tests for gamma gaps and linear model corrected residuals of 1-7 and 2-6 materials

I. Statistical analysis

The linear model was generated using atom identities as categorical variables for the gaps. The residuals therefore encompass the nonlinear components of the atom identity to gap relation. The Kolmogorov-Smirnov tests suggest that the datasets posted in table S1 are from different enough distributions to impede learning. This further suggests low transferability between the rocksalts and zincblendes as well as between oxidation states.

Broken down into individual atoms, the lack of trends within groups is also apparent (Table S2). Anions, in general, have low correlation with each other based on Kolmogorov-Smirnov tests. Notably, F and O are expectedly outliers from the overall distribution. On the cations' side, the trends are more readily available. Residuals are more uniform with each other and thus suggest an overall shared shape. However, the gamma gaps themselves are wildly varying, suggesting that the trends are not consistent in energy, which disrupts the fitting of the base linear model in the first place.

This was noted when trying to project the surfaces from Figure S8 such that the points lined up. While it was possible to align significant regions of the surfaces to each other, it was impossible to align them all. This suggested a fragmented pattern which is difficult to generalize, especially with the now heavily limited dataset of four to five points available.

	Gamma Gap		Residual	
Atom	D	p	D	p
F	0.800	0.00877	1.000	0.00013
Cl	0.533	0.20537	0.600	0.10952
Br	0.533	0.20537	0.533	0.20537
I	0.933	0.00077	0.867	0.00271
O	1.000	0.00013	1.000	0.00013
S	1.000	0.00013	0.467	0.34262
Se	1.000	0.00013	0.400	0.54902
Te	1.000	0.00013	0.667	0.05186
Li	0.313	0.87100	0.375	0.70836
Na	0.750	0.03385	0.313	0.87100
K	0.188	0.99979	0.188	0.99979
Rb	0.250	0.98328	0.250	0.98328
Cs	0.375	0.70836	0.250	0.98328
Be	0.813	0.01445	0.500	0.34757
Mg	1.000	0.00041	0.500	0.34757
Ca	0.438	0.52136	0.313	0.87100
Sr	0.500	0.34757	0.250	0.98328
Ba	0.375	0.70836	0.188	0.99979

TABLE S2. Kolmogorov-Smirnov tests for gamma gap and residuals for individual atoms of zincblende materials within their respective oxidation states.

J. Beyond PBE

The primary motivations behind using PBE-based gaps as the target were cost and presumed complexity. As the majority of bandgaps provided by the Materials Project at time of writing is still based in PBE, computing the gamma gaps in the same way maintained consistency and provided initial geometries, easing computation.[4] It is also presumed that the larger amount of approximations in PBE would reduce the complexity of the resulting function to learn.

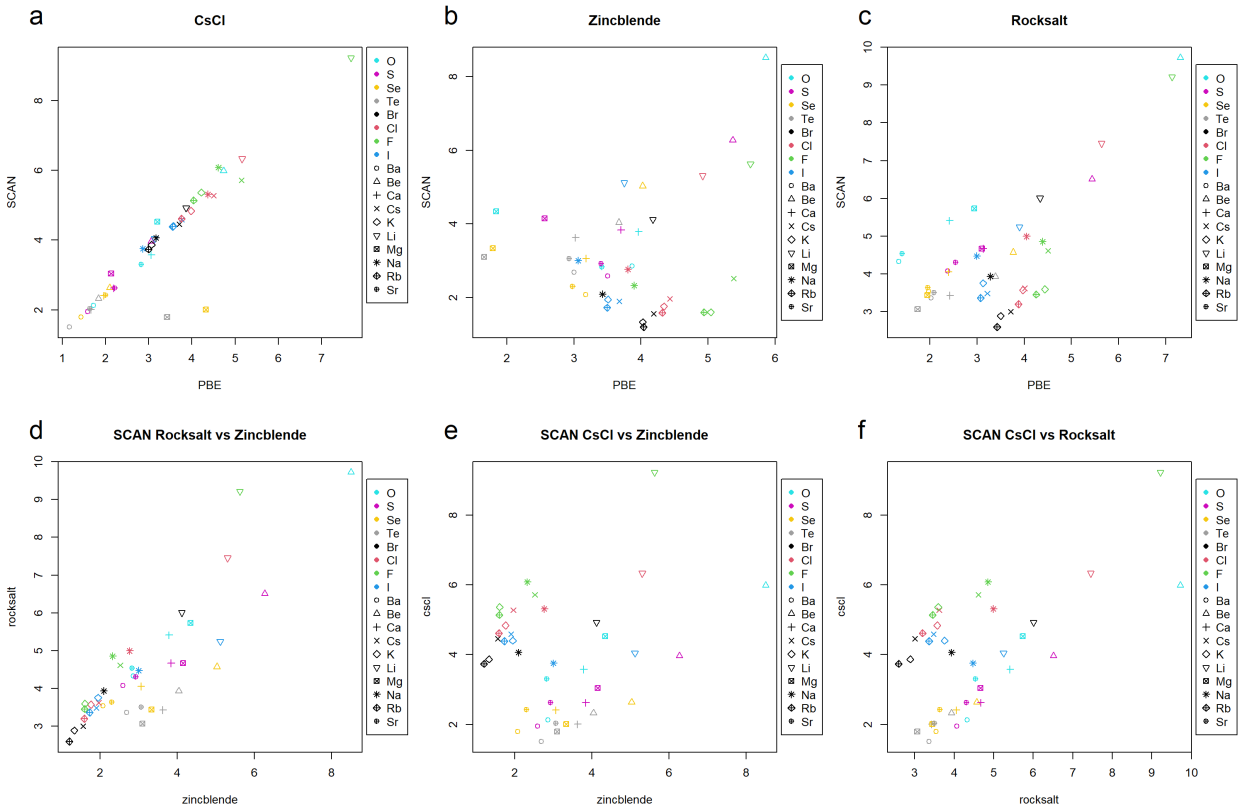

FIG. S17. Comparisons between r2SCAN and PBE gaps for **a,** CsCl, **b,** zincblende, and **c,** rocksalt structures. **d-f,** Additional cross comparisons for r2SCAN on the respective pairs are additionally presented in order.

This, however, does present the potential for more realistic methods to instead smooth the transference between the functions. Comparisons between PBE and r2SCAN show significant differences between the methods for rocksalts and zincblendes, although fairly good agreement for CsCls (Fig. S17. A computation of the materials using r2SCAN and HSE06 were conducted showing closer alignment between the structures but still offer enough differences to impede full transfer between surfaces (Fig. S18-S19).

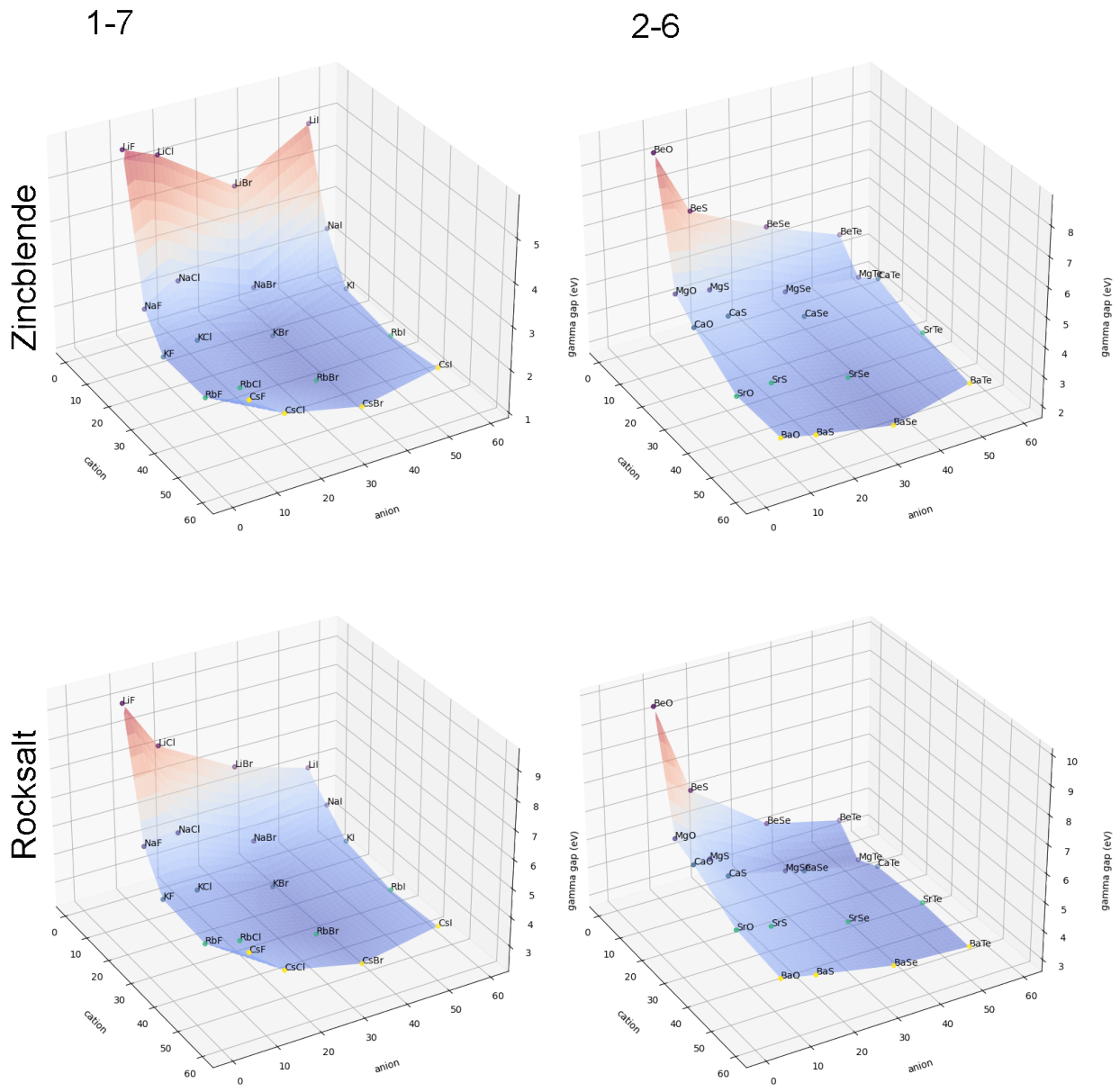

FIG. S18. Visualization of the gamma gaps of the 1-7 and 2-6 rocksalt and zincblende materials across cation and anion identity with r2SCAN.

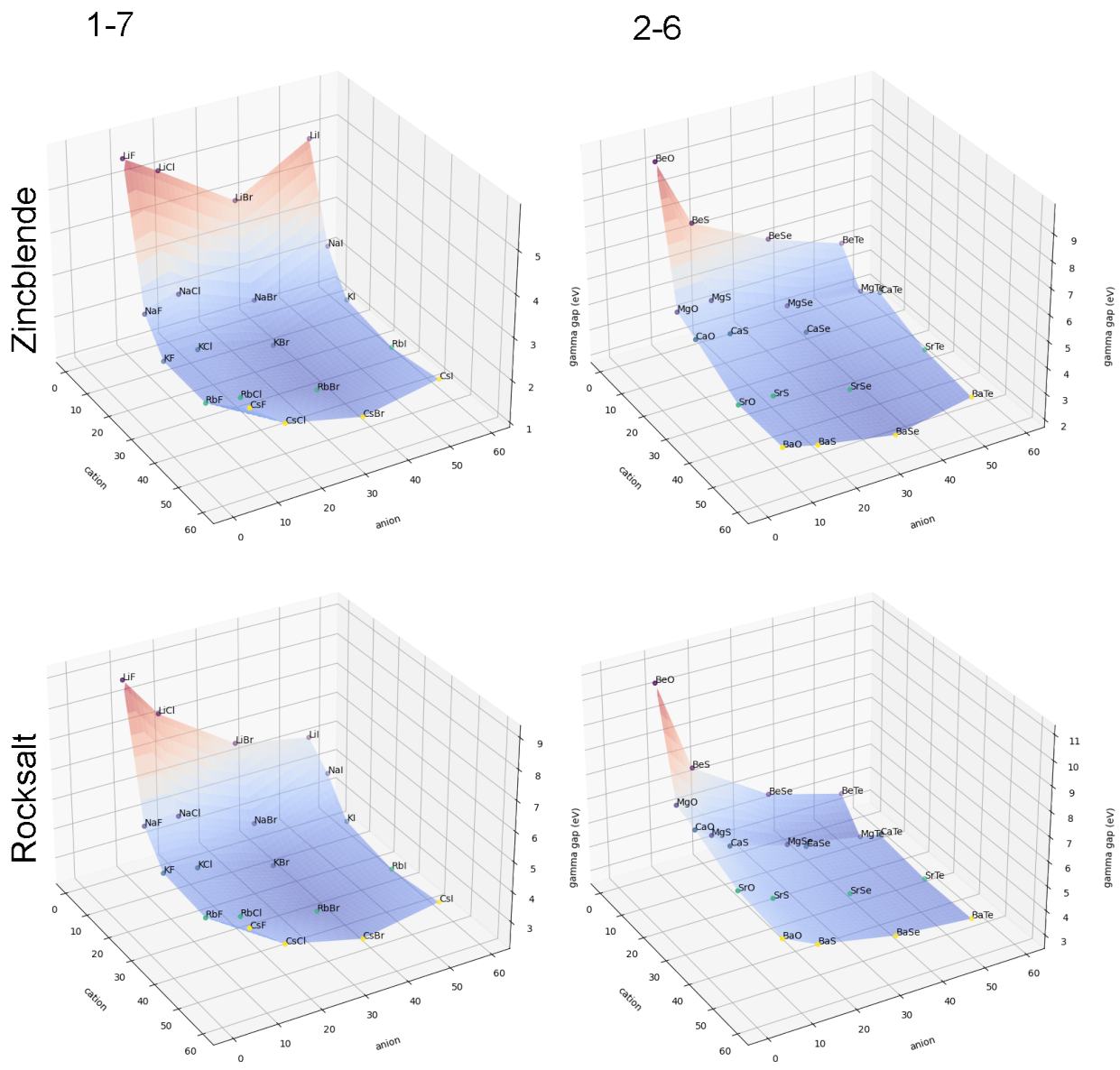

FIG. S19. Visualization of the gamma gaps of the 1-7 and 2-6 rocksalt and zincblende materials across cation and anion identity with r2SCAN.

REFERENCES

[1]T. Xie and J. C. Grossman, "Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties," Phys. Rev. Lett. **120**, 145301 (2018).

[2]A. Dunn, Q. Wang, A. Ganose, D. Dopp, and A. Jain, "Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm," npj Computational Materials **6**, 138 (2020).

[3]R. Ruff, P. Reiser, J. Stühmer, and P. Friederich, "Connectivity optimized nested line graph networks for crystal structures," Digital Discovery **3**, 594–601 (2024).

[4]A. Jain, S. P. Ong, G. Hautier, W. Chen, W. D. Richards, S. Dacek, S. Cholia, D. Gunter, D. Skinner, G. Ceder, and K. a. Persson, "The Materials Project: A materials genome approach to accelerating materials innovation," APL Materials **1**, 011002 (2013).

[5]G. Kresse and D. Joubert, "From ultrasoft pseudopotentials to the projector augmented-wave method," Phys. Rev. B **59**, 1758–1775 (1999).

[6]T. D. Kühne, M. Iannuzzi, M. Del Ben, V. V. Rybkin, P. Seewald, F. Stein, T. Laino, R. Z. Khaliullin, O. Schütt, F. Schiffmann, D. Golze, J. Wilhelm, S. Chulkov, M. H. Bani-Hashemian, V. Weber, U. Borštnik, M. Taillefumier, A. S. Jakobovits, A. Lazzaro, H. Pabst, T. Müller, R. Schade, M. Guidon, S. Andermatt, N. Holmberg, G. K. Schenter, A. Hehn, A. Bussy, F. Belleflamme, G. Tabacchi, A. Glöß, M. Lass, I. Bethune, C. J. Mundy, C. Plessl, M. Watkins, J. VandeVondele, M. Krack, and J. Hutter, "Cp2k: An electronic structure and molecular dynamics software package - quickstep: Efficient and accurate electronic structure calculations," The Journal of Chemical Physics **152**, 194103 (2020).

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

[8]A. H. Larsen, J. J. Mortensen, J. Blomqvist, I. E. Castelli, R. Christensen, M. Dułak, J. Friis, M. N. Groves, B. Hammer, C. Hargus, E. D. Hermes, P. C. Jennings, P. B. Jensen, J. Kermode, J. R. Kitchin, E. L. Kolsbjerg, J. Kubal, K. Kaasbjerg, S. Lysgaard, J. B. Maronsson, T. Maxson, T. Olsen, L. Pastewka, A. Peterson, C. Rostgaard, J. Schiøtz, O. Schütt, M. Strange, K. S. Thygesen, T. Vegge, L. Vilhelmsen, M. Walter, Z. Zeng, and

K. W. Jacobsen, "The atomic simulation environment—a python library for working with atoms," Journal of Physics: Condensed Matter **29**, 273002 (2017).

[9]L. Biewald, "Experiment tracking with weights and biases," (2020), software available from wandb.com.

[10]L. McInnes, J. Healy, N. Saul, and L. Großberger, "Umap: Uniform manifold approximation and projection," Journal of Open Source Software **3**, 861 (2018).

[11]L. Ward, A. Dunn, A. Faghaninia, N. E. Zimmermann, S. Bajaj, Q. Wang, J. Montoya, J. Chen, K. Bystrom, M. Dylla, K. Chard, M. Asta, K. A. Persson, G. J. Snyder, I. Foster, and A. Jain, "Matminer: An open source toolkit for materials data mining," Computational Materials Science **152**, 60–69 (2018).

[12]L. Ward, A. Agrawal, A. Choudhary, and C. Wolverton, "A general-purpose machine learning framework for predicting properties of inorganic materials," npj Computational Materials **2**, 16028 (2016).