Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Illusion of Chemical Continuity Limits Machine Learning Generalization

Oleksandr Voznyy

o.voznyy@utoronto.ca

University of Toronto https://orcid.org/0000-0002-8656-5074

Zhibo Wang

University of Toronto

Alexander Davis

University of Toronto

Rajarshi Dutta

Indian Institute of Technology Kanpur

Alex-Cristian Tomut

University of Toronto

Kanishk Yadav

Ihor Neporozhnii

University of Toronto

Sjoerd Hoogland

University of Toronto https://orcid.org/0000-0002-3099-585X

Article

Keywords:

Posted Date: August 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10463738/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

The Illusion of Chemical Continuity Limits Machine Learning Generalization

Zhibo Wang,[1,2,3] Alexander Davis,[2,3] Rajarshi Dutta,[4] Alex-Cristian Tomut,[1,2] Kanishk Yadav,[5] Ihor Neporozhnii,[2,3] Sjoerd Hoogland,[6,3] and Oleksandr Voznyy[1,2,3,a)]

[1)] *Department of Chemistry, University of Toronto, 10 King's College Road, Toronto, Ontario M5S 3GS, Canada*

[2)] *Department of Physical and Environmental Sciences, University of Toronto Scarborough, 1265 Military Trail, Toronto, Ontario M1C 1A4, Canada*

[3)] *Alliance for AI-Accelerated Materials Discovery*

[4)] *Department of Materials Science and Engineering, Indian Institute of Technology Kanpur, Kalyanpur, Kanpur 208 016, India*

[5)] *Department of Computer Science, Birla Institute of Technology and Science, Pilani, Rajasthan 333 031, India*

[6)] *Electrical and Computer Engineering, University of Toronto, 10 King's College Road, Toronto, Ontario M5S 3GS, Canada*

Abstract

The central dogma of materials informatics is the "Scaling Hypothesis": the assumption that chemical space is a continuous manifold, and that, with sufficient data, deep learning models can approximate any structure-property relationship. While this strategy has achieved near-DFT accuracy for thermodynamic properties like formation energy, prediction accuracy for bandgaps suffers from a fundamental, non-converging error floor ($\sim$0.4 eV). By analyzing a comprehensive dataset of binary semiconductors and state-of-the-art Graph Neural Networks (including coGN and foundational embeddings), this work isolates the root cause: unlike the smooth, transferable landscape of total energy, the electronic bandgap manifold is topologically disjoint. Consequently, increasing dataset size fails to improve predictions, instead resulting in *negative learning*, where increasing the structural diversity of the training set (e.g., adding rocksalt data to zincblende models) degrades prediction accuracy for unseen chemistries. This can be traced back to the discrete quantization of atomic orbitals and the resulting band crossings, which create non-differentiable cusps in the property surface. This work establishes a boundary condition for the "Big Data" paradigm: simple regression on atomic coordinates cannot resolve the intrinsic discontinuities of eigenvalue phenomena. Instead, a more detailed integration of quantum mechanics is required, such as direct entrainment of Hamiltonians as descriptors.

[a)]Corresponding author: o.voznyy@utoronto.ca

1. Introduction

The industrialization of materials science is currently driven by a belief in universal scaling laws. To maximize the success of machine learning (ML), the field has coalesced around the construction of massive, heterogeneous datasets (such as the Materials Project, Open Quantum Materials Database (OQMD), the Open Catalyst Project (OC20), etc.)[1–6] under the assumption that a sufficiently complex neural network can learn the universal functional mapping from crystal structure to physical property.

Despite a decade of architectural innovation and exponential data growth, the root mean squared error (RMSE) for bandgap prediction remains stubbornly high (∼0.4–0.5 eV) with outliers reaching towards 7 eV.[1,7,8] Given that visible light spans ∼1 eV (e.g., blue at ∼2.7 eV, red at ∼1.7 eV), this error renders prediction accuracy inadequate.[9,10]

The prevailing view attributes this stagnation to data scarcity or insufficient model complexity. Ideally, with sufficient data, ML models should be capable of extracting the deeper physical principles governing bandgap formation from surface-level statistical patterns. However, for this to occur, there must exist an underlying systematic trend that ML can leverage.

For bandgaps, this is further complicated by cascading dependencies on bonding and overlaps that modify orbital splittings and order.[11,12]

To probe the root cause of the observed error floor in ML bandgap predictions, this study systematically investigates whether there exist any exploitable trends in the periodic table relating bandgaps and atomic descriptors. To pre-empt complexities associated with orbital crossing, bandgap definitions were standardized to the gamma-point gap to reduce variance. To decouple the impact of structural variation, full combinatorial sets of binary materials were generated and isolated as complete datasets.

By analyzing the topology of chemical space, it is revealed that adding data from different structural domains frequently creates negative transfer, effectively poisoning the model's local accuracy. By systematically eliminating additional confounding factors, this work isolates periodic trends and identifies their limitations in predicting bandgaps. Ultimately, this study reveals that the discontinuity inherent in quantum mechanics, rather than data quantity or model architecture, is what fundamentally restricts ML's ability to generalize bandgap predictions across unseen materials. This suggests that the periodic table representation itself is too coarse-grained to capture the abrupt transitions of the quantum

state.

To encompass the irreducibly quantum nature of bandgaps, the Hamiltonian itself is the best bet to fully describe the complexities of nature of bandgaps.

2. Results

2.1. Disjoint manifolds and local discontinuities

Literature bandgap models report RMSE values around 0.4 eV.[1] When focusing on non-metallic systems only, the RMSE of MatBench's best-performing model, the Connectivity Optimized Nested Line Graph Networks for Crystal Structures (coGN), rises slightly to around 0.5 eV.[1,13]

By quantifying the roughness of the local landscape, it is possible to associate the difficulty of the learned manifold by using the average of a point's nearest neighbours as a naïve predictor of its target value (Fig. 1). For bandgaps, a significant fraction of materials deviate sharply from their immediate neighbours in the latent space.

In contrast, formation energies exhibit higher local continuity, clustering tightly along the diagonal. This topological disparity persists even when using non-learned, training-free

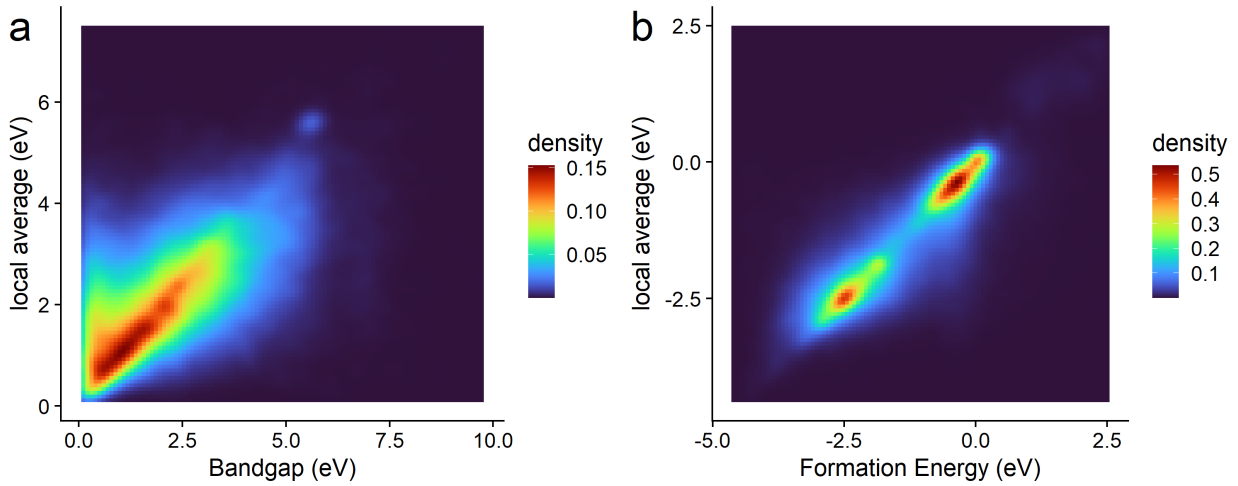


FIG. 1. **Latent space locality analysis. a,** The average value of the 5 nearest neighbours in the coGN latent space is plotted against the actual target. Bandgaps display a high frequency of off-diagonal outliers (spikes near zero), indicating a non-smooth manifold. **b,** In contrast, formation energies exhibit strong local continuity, clustering tightly along the diagonal.

descriptors like Magpie (Fig. S15).[14,15]

This local discontinuity indicates that bandgaps are governed by complex, multi-parametric interactions that defy simple interpolation with current features and model architectures. A classical example are the III-V semiconductors, where bandgap changes from direct to indirect (e.g. upon gradual alloying of GaAs and GaP) resulting in a sharp cusp in bandgap value.[16,17] These factors potentially complicate the effective mapping between structure and bandgap and thus, aimed to be eliminated from the dataset. This highlights the unavoidability of dealing with changing orbital order.

2.2. Structural diversity fails to improve generalization

To eliminate the cusps introduced by indirect bandgaps and k-space anisotropy, the electronic gap taken at the gamma point, the gamma gap, is used in its place. This focuses the learning task on the fundamental energy splitting at the zone center.

To show that the issues experienced in gamma gaps are similar to bandgaps, models were trained on a subset of ∼26000 materials using gamma gaps available from the Materials Project.[4]

Said models (coGN here and CGCNN in the supplement) exhibited a validation RMSE of about 0.7 eV, slightly higher than on fundamental gaps themselves, consistent with gamma gaps being on average noticeably larger (Fig. 2a). Two consistently challenging cases, MgO and NaF zincblendes (known to break common trends in binary materials), were used as a leave-one-out test set, increasing the RMSE to ∼1.0 eV (Fig. 2b).

To assess whether this difficulty reflects data imbalance rather than intrinsic out-of-distribution behavior, coGN was additionally trained on a zincblende-only dataset comprising 74 training structures and 8 validation structures (expanded compared to the 38 available from Materials Project). The MgO and NaF zincblendes were retained as a fixed test set, and all train/validation/test splits were held identical across experiments to ensure direct comparability. Averaged over 10 random initializations, the zincblende-only models achieved an RMSE of ∼0.7 eV on the two test compounds.

To test whether completing missing zincblende entries in the gamma gap dataset improves predictive performance, the expanded zincblende dataset was added to the initial gamma gap dataset, while keeping the zincblende train/validation/test splits fixed. Despite the

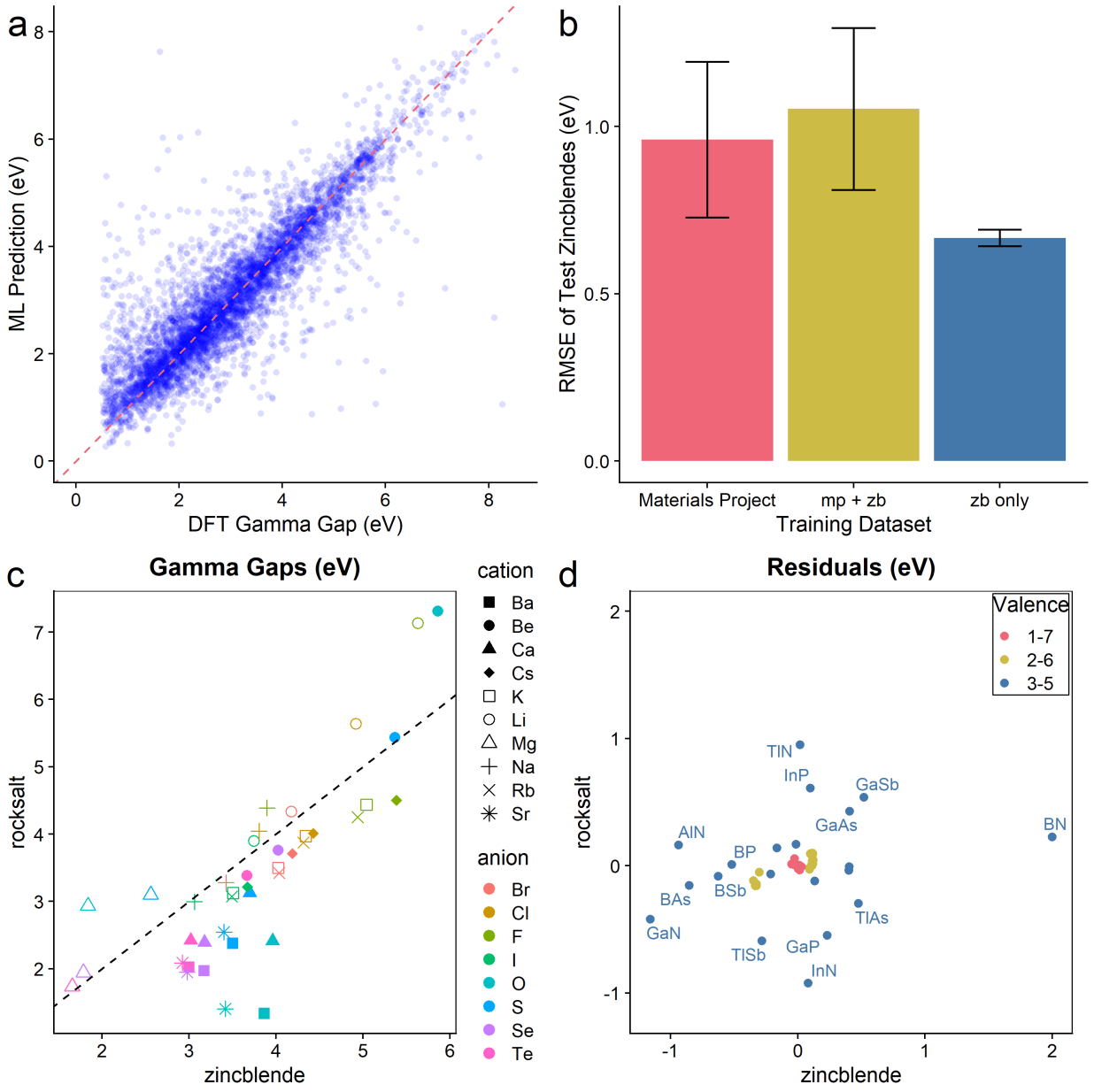


FIG. 2. **Assessment of transferability between crystal structures.** **a,** Validation performance of CoGN trained on Materials Project gamma gaps hovers around 0.7 eV RMSE.[8] **b,** Predictions on two hardest zincblendes (MgO and NaF) trained only on Materials Project gamma gaps (red) and with 90% of zincblendes introduced to the training set (yellow) are outperformed by training solely on zincblendes (blue). **c,** A comparison of rocksalt and zincblende gamma gaps coloured by anion and shaped by cation shows a general correlation with an R^2 of 0.6. **d,** The residuals of linear models based on composition show no trend between the structures with a resulting Pearson's correlation of 0.31.

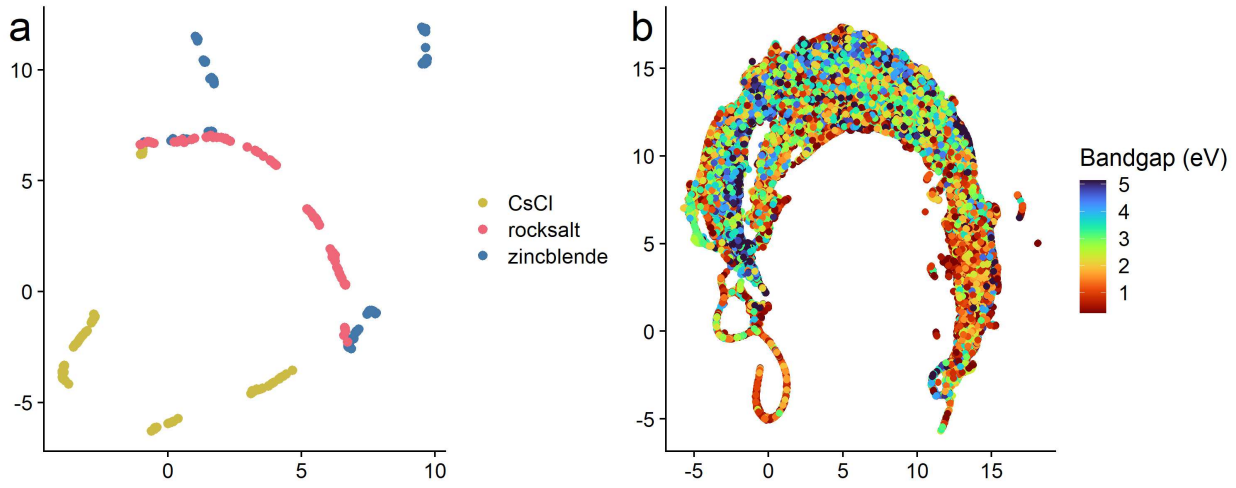


FIG. 3. **Latent space visualizations. a,** UMAP visualization of the coGN latent space embeddings. The three binary structure types, rocksalt (red), zincblende (blue), and CsCl (yellow), occupy disjoint, non-overlapping manifolds, explaining the lack of transferability.[18] **b,** UMAP of the total coGN bandgap space reveals a scattered, uneven prediction manifold.

additional structures, the test RMSE remains ∼1.1 eV, similar to the initial gamma gap dataset. Moreover, the test loss exhibited large variance across random seeds for both the gamma and zincblende added sets, consistent with unstable or near-random predictions typically observed in extrapolation.

To help explain the lack of improvement, the gamma gaps of each zincblende were compared to those of the rocksalt of the same elements. The resulting plot showed a general clustering along the diagonal, suggesting some correlation between the gamma gaps of the two crystal structures (Fig. 2c). However, the deviations from the trend are substantial, with an RMSE of 0.8 eV.

After subtracting a linearly regressed trend based on composition, the residuals show no statistically significant Spearman correlation, with a p-value of 0.31 (Fig. 2d). This confirms that, while intrinsic element properties affect the band gap, interactions between elements in one structure do not make the same contribution in another structure.

The topological origin of this failure is revealed by visualizing the model's latent space using Uniform Manifold Approximation and Projection (UMAP) (Fig. 3a).[18] The embeddings show that rocksalt, zincblende, and CsCl structures occupy distinctly disjoint regions in the latent space. There is no continuous path connecting these manifolds, indicating that

the model treats these structures as chemically dissimilar domains.

Consequently, "Big Data" approaches that aggregate distinct crystal structures effectively force the model to interpolate across a void. As seen in the UMAP of the bandgap latent space (Fig. 3b), this results in a fragmented surface where large gaps are surrounded by smaller ones, generating local outliers (activity cliffs) that defy smooth approximation.[19]

2.3. No trends within groups

Without inter-structural transferability, what remains is for noticeable trends to exist within a single structure (zincblendes, in this case). To evaluate this, the generated zincblende structures were taken and categorized by valence electron count: 1-7 (halides), 2-6 (chalcogenides), and 3-5 (pnictogenides). The 1-7 and 2-6 trends are analyzed here as the most similar, with 3-5 compounds detailed in the supplement (Fig. **??**).

When plotted against atomic identity, the gamma gap surfaces segmented by oxidation state exhibited no overarching pattern between oxidation states (Fig. 4). For example, among third-row cations in zincblendes, Na gaps decreased approximately linearly with increasing anion atomic number (i.e., NaF, NaCl, NaBr, NaI), whereas Mg gaps deviated significantly, with MgO falling significantly below the trend (see red lines). Similar non-monotonicities persist across other element-projected curves as well as in rocksalt structures, consistent with specific electronic localization effects reported in literature.[20]

Statistical analysis further confirmed this lack of intrinsic relationships. Pearson's correlation between the residuals of the two oxidation states was weak ($p = 0.17$), while Spearman's rank correlation was near zero ($\rho = 0.038$, $p = 0.88$), indicating no monotonic relationships. Additionally, a Kolmogorov-Smirnov p-value of 9.5×10^{-6} ($D = 0.75$) confirms that the property distributions of these groups are statistically disjoint.

In sum, even within individual groups, overall trends were absent. Attempts to project a single atom's trend onto another failed to yield consistent mappings. Notably, Na and Mg have Kolmogorov-Smirnov p-values below 0.05 compared to their respective groups (see Table S2 for other atoms).

In contrast, ML-based predictions for total and formation energies were significantly more accurate, suggesting an underlying, learnable trend.[7,8] Comparing these properties to atomic descriptors showed that, unlike bandgaps, total and formation energies exhibited

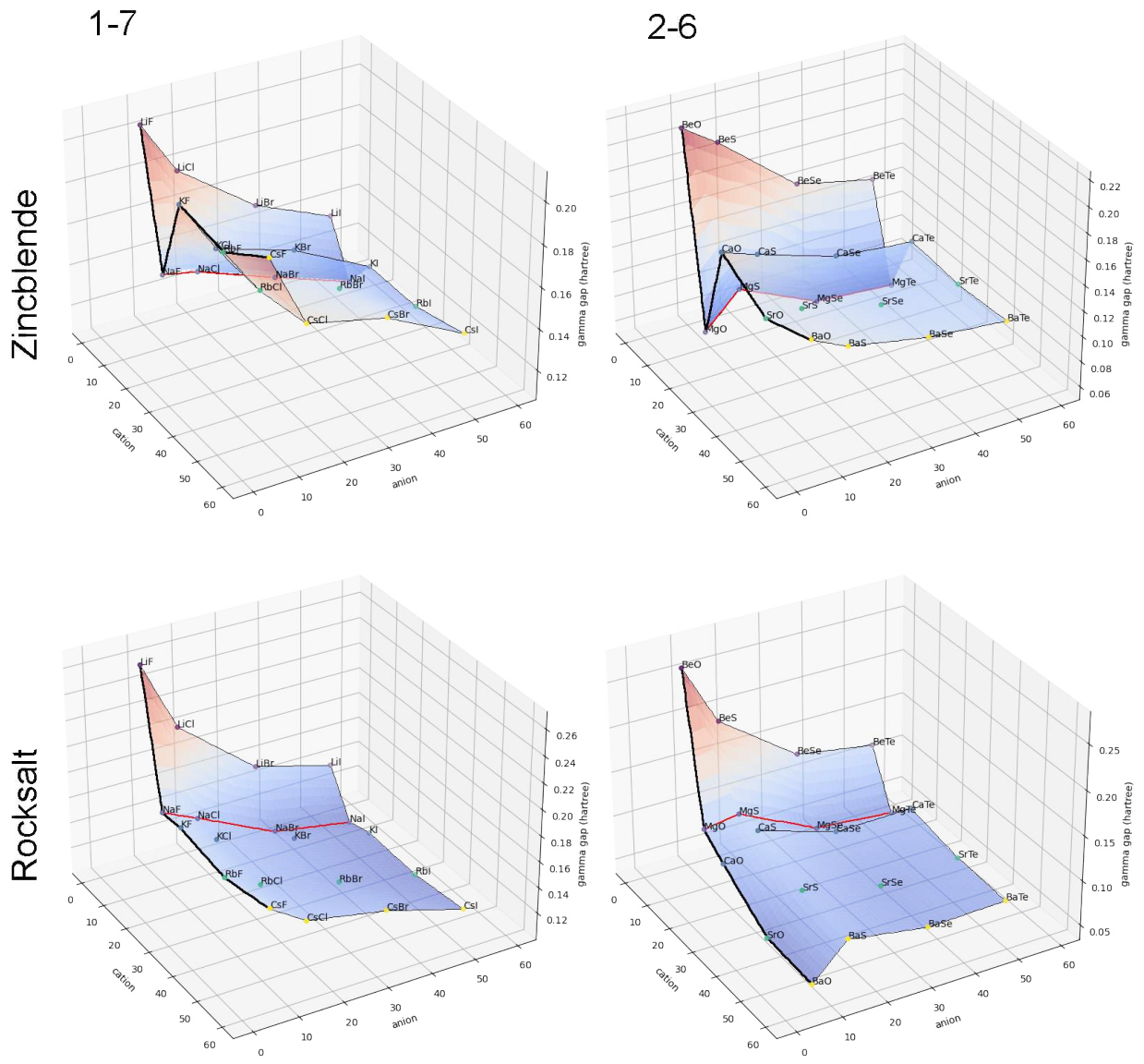


FIG. 4. **Topological fragmentation within crystal families.** Gamma gap surfaces for 1-7 (left) and 2-6 (right) binary compounds in zincblende (top) and rocksalt (bottom) structures. Pairing the compounds by atom row (proxied by atomic number) displays notable deviations in slope by both axes. The oxides and fluorides are outlined in bold, row three cations are outlined in red.

well-defined relationships (Fig. S16).

For these extensive properties, the surfaces for different groups in these cases are similar enough for transference. The total energies are projectable to a shared plane despite group-

wise differences (Fig. S16a). Meanwhile, the formation energy surfaces primarily differ in pitch, with similar patterns across rows (Fig. S16b, c).

2.4. The discreteness of the periodic table undermines machine learning predictions

With no discernible trends among atomic combinations, the remaining hypothesis is that the periodic table itself, the fundamental input space for all materials ML, lacks sufficient continuity to support interpolation for quantum observables.

Among periodic properties, ionization energy is directly relevant to bandgaps, as it proxies the energy of the highest occupied atomic orbital. Plotting these values reveals well-known discontinuities (Fig. 5a), particularly between Groups 2 and 13, where the valence shell transitions from s to p character, and Groups 15 and 16, due to spin-pairing energy penalties.[21–23]

Trends along periods were also irregular. Second-row elements exhibited significantly higher ionization energies than their respective group counterparts, except for alkali metals. In groups 13 and 14, an inflection was observed at the fifth-row elements (In and Sn), which was absent in other groups. These inconsistencies together fragment the periodic table into isolated domains too small to support robust ML predictions (Fig. S11).

2.5. Atomic discontinuities do not interpolate smoothly

To investigate the learnability of the space between these discrete atoms, a simulation of an "alchemical" interpolation between two disjoint species was conducted (Fig. 5b, c). Continuous variation of electron count under a rigid-band approximation from ZnSe to GaSe reveals the stark contrast between total energy versus bandgaps. Whereas total energy follows a smooth, differentiable curve, the electronic gap collapses upon first addition of partial electrons.

Due to the quantized nature of electronic orbitals, the additional fractional electrons force sudden changes in HOMO identity. The Fermi level does not smoothly traverse the bands. Instead, it jumps past the conduction band minimum sharply. Consequently, the periodic table cannot be treated as a continuous vector space for electronic gaps; it is a collection of

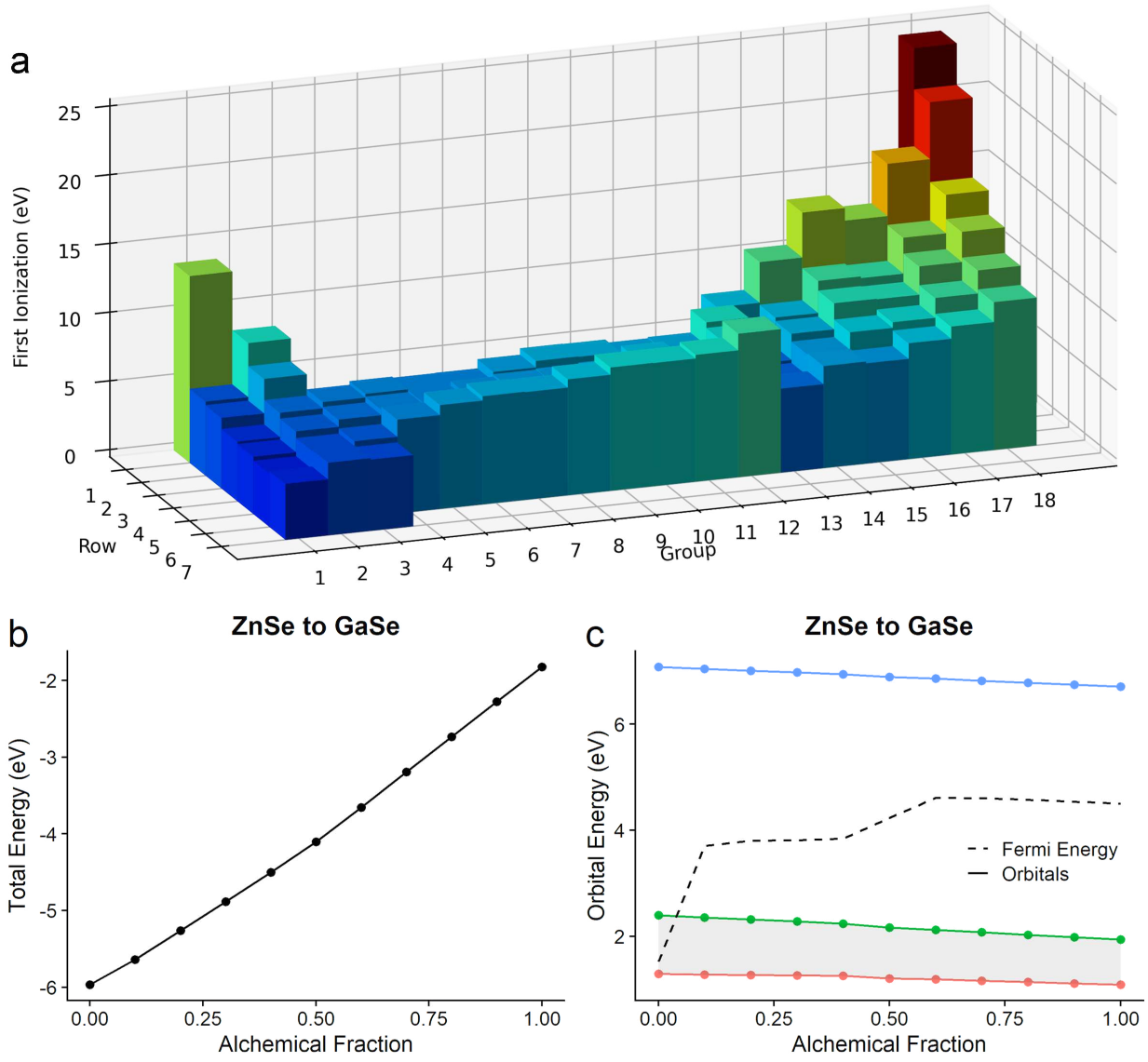


FIG. 5. **The intrinsic topological discreteness of the periodic table. a,** The first ionization energies of the main-group periodic table. The trends are fragmented amongst the individual groups leaving little interpolative ability from one group to another. **b, c,** *Ab initio* simulation of the alchemical interpolation between two such discrete states (ZnSe to GaSe). **b,** While thermodynamic properties like total energy follow a smooth, learnable manifold, **c,** the gamma gap (shaded) exhibits a sharp discontinuity as the Fermi energy (dashed) traverses over the former conduction band at first addition.

discrete, disjoint manifolds.

3. Discussion

Fundamentally, the high predictive loss in ML-based bandgap models arises from the lack of transferable trends across multiple different axes. Structural complexity introduces variations that outweigh the benefits of increased dataset size, resulting in negative learning. This indicates that each crystal structure represents a distinct chemical environment with minimal extrapolation potential, provided that the underlying quantum mechanical rules of orbital interactions have not been learned.

While a weak correlation exists, as seen in Figure 2c, the observed quasi-linear relationships are insufficient for generalization. This failure is most prominent when the linear dependence on composition is removed (Fig. 2d) and in the projections of the embedded structures (Fig. 3a). This fragmentation prevents knowledge learned from the geometry of rocksalt structures to smoothly interpolating to zincblendes, as no continuous path connects them in the latent space.

The issues presented are also independent of model architecture (Fig. S6). CoGN, which handles angular information better, has not helped to resolve it when compared to older models (Fig. S4). The fact that state-of-the-art architectures hit the same wall confirms the limitation lies in the data topology, not the model complexity.

These observations contradict the common assumption that increasing the dataset size automatically improves model performance for bandgaps.[24] Instead, incorporating additional structures introduce noise rather than transferable signal (Fig. 2, S6b). However, when the underlying data manifolds are effectively disjoint, the model is forced to average over incompatible structure–property relationships, leading to a regression-to-the-mean behaviour that obscures the sharp features characteristic of the target class. In other words, the ML model did not even try to learn the physics of orbital interactions.

This limitation persists even within a single crystal structure class. The consistent difficulty in predicting MgO and NaF zincblendes can be understood by examining their local chemical neighborhoods in the projected feature space (Fig. 4). Both compounds exhibit anomalously low band gaps relative to nearby zincblende trends, making them intrinsically hard to interpolate. By contrast, their rocksalt counterparts follow smoother, more uniform

trends. As a result, MgO and NaF behave as effective outliers even within the zincblende manifold, contributing disproportionately to model error.

Visualizing the roughness of the embedded manifold, it is clear that bandgaps lack the consistency needed for prediction from nearest neighbours. This is in contrast with easier properties where the distance is much less severe (Fig. 1).

Bandgaps result from complex quantum mechanical interactions involving orbital overlap, hybridization, and electronic structure. While atomic orbital energies are straightforward to extract from *ab initio* calculations, the interactions governing bandgap formation introduce significant non-linearity and structure dependence.

The ultimate constraint, however, is the finite and discrete domain of the periodic table itself. While modern databases contain massive numbers of crystal structures, the effective dimensionality of the problem is restricted to the $\sim$80 stable elements that comprise them. The inability of ML to comprehend the quantum nature of orbitals (of half-filled shells and ℓ) despite knowledge of the number of valence electrons and groups functionally subdivides the periodic table into even smaller domains. Such domains heavily limit the available training, especially considering the non-transference between crystal structures.

3.1. From property interpolation to Hamiltonian learning

Overcoming the predictive ceiling in ML-based bandgap models requires a fundamental shift in methodology. The alchemical transmutation analysis reveals that the failure of current models is not due to data scarcity, but target selection (Fig. 5). While bandgaps exhibit non-differentiable trends that prevent interpolation, the underlying orbital eigenvalues evolve more smoothly. In effect, the semi-conducting state is a feature evolving from the Hamiltonian, not a smooth function of atomic coordinates.

Consequently, future efforts must pivot from direct property prediction towards learning the underlying quantum mathematics. This likely requires repurposing the graph representations from molecular structures for the Hamiltonian matrix elements instead. A model capable of predicting the continuous Hamiltonian matrix elements or electron density may better preserve the smooth physics of orbital interactions, allowing the discrete bandgap to be recovered subsequently.[25,26]

Until ML models integrate explicit quantum mechanical principles, bandgap predictions

will remain constrained by the inherent discontinuities of chemical space.

4. Conclusions

Ab initio computations use physics to derive properties; ML uses data to derive trends. For ML to replicate *ab initio* accuracies, the physical property in question must exhibit an emergent trend that can be smoothly mapped to the input descriptors. Bandgaps are not such a property. There are no continuous functions that map atom identities to orbital splittings directly, the energy level crossings force piecewise assemblies that hinder mapping and orbital order changes within k-space.

This predictive barrier is rooted in the intrinsic topological fragmentation of the periodic table. Instead of a unified foundation built upon 118 atoms in a single table, one must rely on 18 groups of piecewise trends. Additional training from one such disjoint group does not meaningfully improve the learning of another, it instead leads to *negative learning*. Consequently, the independent structures contain insufficient data points to reconstruct a representative function that can generalize to unseen materials and the complexities of their bandgaps.

5. Methods

5.1. Database preparation

Crystalline materials exhibit periodicity, leading to variations in orbital energy ordering across wave vectors (k). This introduces distinctions between direct and indirect bandgaps. As the highest occupied and lowest unoccupied molecular orbitals (HOMO and LUMO) can vary across k-points, bandgap definitions become piecewise functions. The prediction task was simplified by using bandgaps at the gamma point to remove this dependence.

The initial dataset was obtained from the Materials Project (snapshot: June 2023) using API queries.[4] Bandstructures for a subset of materials (26331) were also retrieved for further analysis of bandgap characteristics and gamma point gaps. For reference model outputs, datasets were extracted from Matbench for use on benchmarked models alongside in-house datasets.[1] For comparisons with gamma gaps, metals were filtered out of the bandgaps datasets for gaps below 0.1 eV. A notable deficiency in the dataset was the ab-

sence of zincblende binary compounds, particularly among alkali and alkaline earth elements, resulting in incomplete coverage of this structure type.

To address these deficiencies, missing zincblende structures were computed and optimized using density functional theory (DFT) for comparative analysis. Additionally, rocksalt and CsCl structures were included for further evaluation of structure-dependent trends. Structures were optimized in the Vienna Ab initio Simulation Package (VASP) with the Perdew-Burke-Ernzerhof (PBE) generalized gradient approximation functional and projector augmented wave (PAW) pseudopotentials.[27–31] Optimizations were performed in two steps, first with Gaussian smearing and then with tetrahedron smearing.

For further characterization, projected density of states (PDOS) calculations were performed in CP2K, which uses explicit atomic orbital projections.[32–47] Atomic orbitals were similarly computed with PBE/DZVP-MOLOPT-SR-GTH level of theory with GTH-PBE pseudopotentials to maintain congruency between atomic and crystal data. These orbitals were used for visualization of gap topologies and orbital conributions.

To extract atomic descriptors, properties such as electronegativity, ionization energy, and polarizability were retrieved from Pymatgen[22] and the Atomic Simulation Environment (ASE).[23] Additional properties, including hardness, were sourced from external literature where necessary.[48,49] These properties were compiled into JSON lookup tables, facilitating conversion into numerical feature vectors. Discrete variables were encoded using one-hot representations, while continuous variables were normalized.

To probe the intrinsic topology and continuity of the electronic manifold, an *ab initio* simulation modelling the transition between ZnSe to GaSe was performed. Computations were conducted as stated prior with VASP and PBE with PAW pseudopotentials, incrementing the total electron count by increments of 0.1.[27–31] A compensating homogeneous background charge (jellium) was automatically applied to maintain cell neutrality, effectively simulating a fractional increase in nuclear charge. To decouple electronic topological features from structural distortion, the ionic positions were fixed to the relaxed ground-state geometry of ZnSe.

5.2. ML modelling

Custom, non-graphical ML models were implemented in Python using PyTorch.[50] The Crystal Graph Convolutional Neural Network (CGCNN) was employed as the reference architecture, maintaining standard feature conventions.[8] Inputs were structured into atomic and structural representations and then combined to form atomic feature vectors specific to each crystal system. Additional comparisons were conducted using Connectivity Optimized Nested Line Graph Networks for Crystal Structures (CoGN) for a more modern reference.[1,13]

The crystal data were formatted to contain compound features, the target parameter of choice, crystal structure identifier (bonding neighbours), and structure-specific properties that were not encodable by atomic property combinations. This generates a final vector that concatenates the cation, anion, and additional crystal properties. A detailed breakdown is included in the supporting information. CGCNN was used as per documentation with extracted hyperparameters from pre-trained models.

Training and validation datasets were generated via random sampling, with fixed random seeds ensuring reproducibility. Across comparisons involving multiple crystal structure types, a consistent set of the most challenging zincblende materials was retained exclusively for testing, ensuring that the validation and test partitions remained identical between experiments. For architecture comparisons and preliminary identification of difficult structures, the small dataset size necessitated omitting a dedicated test set; these data points were instead blended into the training/validation split.

Hyperparameter exploration was performed using Weights & Biases to obtain coarse estimates of layer widths and layer counts.[51] Automated tuning did not always provide stable hyperparamenters leaving exact learning rates and weight decays to human input. Dataset comparisons were conducted using consistent hyperparameter settings to ensure fair evaluations. Literature models were initialized with published best-performing configurations or default settings rather than re-tuned from scratch.

For comparing structures, a subset of extracted gamma gaps from Materials Project was used to train a baseline model. This was split into an 80-20% training and validation set of 26331 gamma gaps. For comparison, another model was trained with the same training dataset along with 74 additional binary zincblendes. A third model of only binary zincblendes was also constructed with the same 74 zincblendes. The three models were

validated against the same set of 8 randomly selected zincblendes with the two hardest materials as an unseen test set (MgO and NaF). Comparative models were run with identical splits and hyperparameters to maintain unity. 10 random initial seeds were conducted and standard errors visualized for comparisons of predictive stability.

To eliminate the possible inability for graph neural networks to fully encode the necessary geometric information, a structurally identical subset of materials was used for architectural contributions to accuracy. This was done through the systematic isolation of model components, generating five architectures as follows: a forward-feeding set of linear layers, said forward-feeding layers with the atom vectors convolved together, the convolved atoms pooled together, a static graph without pooling, and a pooled graph (the entire graph convolutional neural network intact).

Visualizations of the coGN latent space were conducted by training models with full training coverage until no further improvements were found.[1] Bandgap analysis was focused on non-metallic (¿ 0.1 eV) materials for clarity. The resultant models were then processed via Uniform Manifold Approximation and Projections (UMAP) and Principal Component Analysis (PCA) to analyze the relative roughness of the underlying properties.[18,19] Additional visualizations of the local manifold were also performed using Matminer features for comparisons between featurizers.

5.3. Statistical trend analysis

Even with a fixed set of elements, datasets can be expanded by including crystal structures with different space groups. Therefore, the transferability of learning across space groups was assessed by comparing gamma gaps for zincblende and rocksalt crystal structures with the same composition. Gamma gap distributions for zincblende and rocksalt materials were examined for similarities in morphology besides packing/neighbouring atoms, and their elemental combinations were plotted to identify periodic trends. Linear models of the gaps were used to adjust for the contributions from cation and anion identities. The residuals, therefore, represent an adjusted value for the gap relative to the expectations based on composition.

To further assess the presence of predictive trends within zincblende structures, statistical tests were performed on gamma gap residuals obtained. Structures were paired by anion and

cation row using the residuals from a linear model to eliminate the expected contributions by element position as before. Residuals were analyzed to capture deviations from expected atomic contributions.

Pearson and Spearman correlation coefficients were computed between the residuals for 1-7 and 2-6 compounds to evaluate potential linear and monotonic relationships. A linear regression model was applied to determine whether one surface provided significant predictive power for the other, with R^2 values used to evaluate the proportion of variance explained. Additionally, two-sample Kolmogorov-Smirnov (KS) tests were performed to assess whether the oxidation states followed similar distributions, with the p-value quantifying the significance of differences between their cumulative distribution functions.

Atomic properties from the periodic table were collated and projected by group and row. Groupwise patterns were compared with each other by projecting onto group and row. Similar models to the previous were constructed taking group and period as the parameters to obtain nonlinear relations.

6. Author contributions

Zhibo Wang: conceptualization, methodology, programming, modelling, formal analysis, investigation, data curation, writing – original draft, writing – review and editing, visualization, project administration. Alexander Davis: statistics, investigation. Rajarshi Dutta: hyperparameter tuning. Alex-Cristian Tomut: additional DFT optimization. Kanishk Yadav and Ihor Neporozhnii: additional models, coding. Sjoerd Hoogland: funding acquisition, project administration. Oleksandr Voznyy: supervision, conceptualization, resources, funding acquisition, writing – review and editing, project administration.

7. Conflicts of interest

There are no conflicts to declare.

8. Data availability

Data for this article, including code, models, training data, Materials Project identification numbers, and input structures for reproduction are available on GitHub at

https://github.com/Voznyy-Clean-Energy-Lab-UToronto/ABLinear.[52]

9. Acknowledgements

The authors acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) Discover Grant (2024-05613) and the Alliance for AI-Accelerated Materials Discovery (A3MD) for financing this project. This work was enabled in part due to the resources and services provided by Compute/Calcul Canada (computecanada.ca) and computations were performed on the Niagara supercomputer at the SciNet HPC Consortium.[53,54] SciNet is funded by the Canada Foundation for Innovation; the Government of Ontario; Ontario Research Fund - Research Excellence; and the University of Toronto.

References

[1]Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: the matbench test set and automatminer reference algorithm. *npj Computational Materials* **6**, 138 (2020). URL `https://doi.org/10.1038/s41524-020-00406-3`.

[2]Merchant, A. *et al.* Scaling deep learning for materials discovery. *Nature* **624**, 80–85 (2023). URL `https://doi.org/10.1038/s41586-023-06735-9`.

[3]Chen, C. & Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science* **2**, 718–728 (2022). URL `https://doi.org/10.1038/s43588-022-00349-3`.

[4]Jain, A. *et al.* The Materials Project: A materials genome approach to accelerating materials innovation. *APL Materials* **1**, 011002 (2013).

[5]Saal, J. E., Kirklin, S., Aykol, M., Meredig, B. & Wolverton, C. Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd). *JOM* **65**, 1501–1509 (2013). URL `https://doi.org/10.1007/s11837-013-0755-4`.

[6]Kirklin, S. *et al.* The open quantum materials database (oqmd): assessing the accuracy of dft formation energies. *npj Computational Materials* **1**, 15010 (2015). URL `https://doi.org/10.1038/npjcompumats.2015.10`.

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

[8]Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018). URL https://link.aps.org/doi/10.1103/PhysRevLett.120.145301.

[9]Chen, C. *et al.* A critical review of machine learning of energy materials. *Advanced Energy Materials* **10**, 1903242 (2020). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/aenm.201903242. https://onlinelibrary.wiley.com/doi/pdf/10.1002/aenm.201903242.

[10]Yılmaz, B. & Yıldırım, R. Critical review of machine learning applications in perovskite solar research. *Nano Energy* **80**, 105546 (2021). URL https://www.sciencedirect.com/science/article/pii/S2211285520311204.

[11]Jang, D., Han, Y., Baek, S. & Kim, J. Theoretical comparison of the energies and wave functions of the electron and hole states between cdse- and inp-based core/shell/shell quantum dots: effect of the bandgap energy of the core material on the emission spectrum. *Opt. Mater. Express* **9**, 1257–1270 (2019). URL https://opg.optica.org/ome/abstract.cfm?URI=ome-9-3-1257.

[12]Allen, C. R., Liu, S. & Zhang, Y.-H. Ii-vi/iii-v integration for next generation multijunction solar cells. *2011 37th IEEE Photovoltaic Specialists Conference* 000225–000227 (2011). URL https://api.semanticscholar.org/CorpusID:43538076.

[13]Ruff, R., Reiser, P., Stühmer, J. & Friederich, P. Connectivity optimized nested line graph networks for crystal structures. *Digital Discovery* **3**, 594–601 (2024). URL http://dx.doi.org/10.1039/D4DD00018H.

[14]Ward, L. *et al.* Matminer: An open source toolkit for materials data mining. *Computational Materials Science* **152**, 60–69 (2018). URL https://www.sciencedirect.com/science/article/pii/S0927025618303252.

[15]Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials* **2**, 16028 (2016). URL https://doi.org/10.1038/npjcompumats.2016.28.

[16]González, M. *et al.* Modeling, design and experimental results for high efficiency multijunction solar cells lattice matched to InP. In Freundlich, A. & Guillemoles, J.-F. (eds.)

Physics, Simulation, and Photonic Engineering of Photovoltaic Devices III, vol. 8981, 898117. International Society for Optics and Photonics (SPIE, 2014). URL https://doi.org/10.1117/12.2041289.

[17]Vurgaftman, I., Meyer, J. R. & Ram-Mohan, L. R. Band parameters for iii–v compound semiconductors and their alloys. *Journal of Applied Physics* **89**, 5815–5875 (2001). URL https://doi.org/10.1063/1.1368156.

[18]McInnes, L., Healy, J., Saul, N. & Großberger, L. Umap: Uniform manifold approximation and projection. *Journal of Open Source Software* **3**, 861 (2018). URL https://doi.org/10.21105/joss.00861.

[19]Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[20]Yu, J., Zhang, M., Zhang, Z., Wang, S. & Wu, Y. Hybrid-functional calculations of electronic structure and phase stability of mo (m = zn, cd, be, mg, ca, sr, ba) and related ternary alloy mxzn1-xo. *RSC Adv.* **9**, 8507–8514 (2019). URL http://dx.doi.org/10.1039/C9RA00362B.

[21]Linstrom, P. J. & Mallard, G. W. *NIST Chemistry WebBook, NIST Standard Reference Database 69* (National Institute of Standards and Technology, 1997). URL http://webbook.nist.gov/chemistry/.

[22]Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Computational Materials Science* **68**, 314–319 (2013).

[23]Larsen, A. H. *et al.* The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter* **29**, 273002 (2017). URL https://dx.doi.org/10.1088/1361-648X/aa680e.

[24]Li, K. *et al.* A call for caution in the era of ai-accelerated materials science. *Matter* **6**, 4116–4117 (2023). URL https://www.sciencedirect.com/science/article/pii/S2590238523005283.

[25]Bai, H. *et al.* Graph neural network for hamiltonian-based material property prediction. *Neural Computing and Applications* **34**, 4625–4632 (2022). URL https://doi.org/10.1007/s00521-021-06616-0.

[26]Li, Y., Zhang, X., Liu, M. & Shen, L. A critical review of machine learning interatomic potentials and hamiltonian. *Journal of Surveillance, Security and Safety* **5**, 43 (2025). URL https://www.oaepublish.com/articles/jmi.2025.17.

[27]Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996). URL https://link.aps.org/doi/10.1103/PhysRevB.54.11169.

[28]Kresse, G. & Hafner, J. Ab initio molecular dynamics for liquid metals. *Phys. Rev. B* **47**, 558–561 (1993). URL https://link.aps.org/doi/10.1103/PhysRevB.47.558.

[29]Kresse, G. & Joubert, D. From ultrasoft pseudopotentials to the projector augmented-wave method. *Phys. Rev. B* **59**, 1758–1775 (1999). URL https://link.aps.org/doi/10.1103/PhysRevB.59.1758.

[30]Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994). URL https://link.aps.org/doi/10.1103/PhysRevB.50.17953.

[31]Wang, V., Xu, N., Liu, J.-C., Tang, G. & Geng, W.-T. Vaspkit: A user-friendly interface facilitating high-throughput computing and analysis using vasp code. *Computer Physics Communications* **267**, 108033 (2021).

[32]Kühne, T. D. *et al.* Cp2k: An electronic structure and molecular dynamics software package - quickstep: Efficient and accurate electronic structure calculations. *The Journal of Chemical Physics* **152**, 194103 (2020). URL https://doi.org/10.1063/5.0007045.

[33]Schütt, O., Messmer, P., Hutter, J. & VandeVondele, J. *GPU-Accelerated Sparse Matrix–Matrix Multiplication for Linear Scaling Density Functional Theory*, chap. 8, 173–190 (John Wiley & Sons, Ltd, 2016). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118670712.ch8. https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118670712.ch8.

[34]Borštnik, U., VandeVondele, J., Weber, V. & Hutter, J. Sparse matrix multiplication: The distributed block-compressed sparse row library. *Parallel Computing* **40**, 47–58 (2014). URL https://www.sciencedirect.com/science/article/pii/S0167819114000428.

[35]Hutter, J., Iannuzzi, M., Schiffmann, F. & VandeVondele, J. cp2k: atomistic simulations of condensed matter systems. *WIREs Computational Molecular Science* **4**, 15–25 (2013). URL https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/wcms.1159.

[36]Grimme, S., Ehrlich, S. & Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *Journal of Computational Chemistry* **32**, 1456–1465 (2011). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.21759.

[37]Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (dft-d) for the 94 elements h-

pu. *The Journal of Chemical Physics* **132**, 154104 (2010). URL https://doi.org/10.1063/1.3382344.

[38]VandeVondele, J. & Hutter, J. Gaussian basis sets for accurate calculations on molecular systems in gas and condensed phases. *The Journal of Chemical Physics* **127**, 114105 (2007). URL https://doi.org/10.1063/1.2770708.

[39]Krack, M. Pseudopotentials for h to kr optimized for gradient-corrected exchange-correlation functionals. *Theoretical Chemistry Accounts* **114**, 145–152 (2005). URL https://doi.org/10.1007/s00214-005-0655-y.

[40]VandeVondele, J. *et al.* Quickstep: Fast and accurate density functional calculations using a mixed gaussian and plane waves approach. *Computer Physics Communications* **167**, 103–128 (2005). URL https://www.sciencedirect.com/science/article/pii/S0010465505000615.

[41]Frigo, M. & Johnson, S. The design and implementation of fftw3. *Proceedings of the IEEE* **93**, 216–231 (2005).

[42]Kolafa, J. Time-reversible always stable predictor–corrector method for molecular dynamics of polarizable molecules. *Journal of Computational Chemistry* **25**, 335–342 (2004). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/jcc.10385.

[43]VandeVondele, J. & Hutter, J. An efficient orbital transformation method for electronic structure calculations. *The Journal of Chemical Physics* **118**, 4365–4369 (2003). URL https://doi.org/10.1063/1.1543154.

[44]Hartwigsen, C., Goedecker, S. & Hutter, J. Relativistic separable dual-space gaussian pseudopotentials from h to rn. *Phys. Rev. B* **58**, 3641–3662 (1998). URL https://link.aps.org/doi/10.1103/PhysRevB.58.3641.

[45]LIPPERT, B. G., HUTTER, J. & PARRINELLO, M. A hybrid gaussian and plane wave density functional scheme. *Molecular Physics* **92**, 477–488 (1997). URL https://doi.org/10.1080/002689797170220.

[46]Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996). URL https://link.aps.org/doi/10.1103/PhysRevLett.77.3865.

[47]Goedecker, S., Teter, M. & Hutter, J. Separable dual-space gaussian pseudopotentials. *Phys. Rev. B* **54**, 1703–1710 (1996). URL https://link.aps.org/doi/10.1103/PhysRevB.54.1703.

[48]Dong, X., Oganov, A. R., Cui, H., Zhou, X.-F. & Wang, H.-T. Electronegativity and chemical hardness of elements under pressure. *Proceedings of the National Academy of Sciences* **119**, e2117416119 (2022). URL https://www.pnas.org/doi/abs/10.1073/pnas.2117416119. https://www.pnas.org/doi/pdf/10.1073/pnas.2117416119.

[49]Schwerdtfeger, P. & Nagle, J. K. 2018 table of static dipole polarizabilities of the neutral elements in the periodic table*. *Molecular Physics* **117**, 1200–1225 (2019).

[50]Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*, 8024–8035 (Curran Associates, Inc., 2019). URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[51]Biewald, L. Experiment tracking with weights and biases (2020). URL https://www.wandb.com/. Software available from wandb.com.

[52]URL https://github.com/Voznyy-Clean-Energy-Lab-UToronto/ABLinear. Modularized two-atom network: https://github.com/Voznyy-Clean-Energy-Lab-UToronto/ABLinear.

[53]Ponce, M. *et al.* Deploying a top-100 supercomputer for large parallel workloads: The niagara supercomputer. In *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning)*, PEARC '19 (Association for Computing Machinery, New York, NY, USA, 2019). URL https://doi.org/10.1145/3332186.3332195.

[54]Loken, C. *et al.* SciNet: Lessons learned from building a power-efficient top-20 system and data centre. *Journal of Physics: Conference Series* **256**, 012026 (2010). URL https://doi.org/10.1088/1742-6596/256/1/012026.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TrendlessPeriodicTableSI.pdf