26 May 2026

Transfer learning on universal interatomic potential embeddings improves generalization in structure-property defect models

Matthew D. Witman, Sebastian Pujet, Andrew J. E. Rowberg, Christopher Sutton, Joel B. Varley, Stephan Lany, Robert B. Wexler

Abstract

Oxygen vacancy formation energies govern the performance of metal oxides across energy conversion, catalysis, and electronics, yet predicting them accurately without density functional theory calculations for novel chemical systems remains challenging. Here, we show that frozen 256-dimensional embeddings extracted from a pretrained MACE universal machine-learning interatomic potential encode information sufficient to predict vacancy formation energies obtained from density functional theory without requiring supercell construction, vacancy creation, or geometry optimization. A lightweight multilayer perceptron trained on these embeddings (MACE-dGNN) achieves an element-wise cross-validation mean absolute error of 0.38 eV, halving the error of a baseline graph neural network trained from scratch (0.72 eV) and offering a modest advantage over direct MACE relaxation calculations (0.51 eV). Because the embedding approach decouples representation from the prediction task and generalizes well with limited data, it will (1) extend naturally to fine-tuning and prediction of properties (e.g., charged defects) that are inaccessible to generic energy-and force-output interatomic potentials and (2) be trainable on necessarily small datasets, potentially enabling the use of more expensive but accurate tools for generating training data (e.g., hybrid functional calculations) for which training or fine-tuning of interatomic potentials will be difficult. We demonstrate practical impact by revisiting a prior screening of thermochemical water-splitting materials, where improved generalization alters #45 % of candidate classifications.

 https://doi.org/10.26434/chemrxiv.15003938/v1

Transfer learning on universal interatomic potential embeddings improves generalization in structure-property defect models

Matthew D. Witman,[1,*] Sebastian Pujet,[2] Andrew J. E. Rowberg,[3] Christopher Sutton,[4,5] Joel B. Varley,[3] Stephan Lany,[6] and Robert B. Wexler[7,†]

[1]*Sandia National Laboratories, Livermore, CA 94550, USA*

[2]*Department of Energy, Environmental & Chemical Engineering, Washington University in St. Louis, St. Louis, MO 63130, USA*

[3]*Lawrence Livermore National Laboratory, Livermore, CA 94550, USA*

[4]*Department of Materials Science and Engineering, University of Toronto, Toronto, ON M5S 3E4, Canada*

[5]*Department of Chemistry and Biochemistry, University of South Carolina, Columbia, SC 29208, USA*

[6]*National Laboratory of the Rockies, Golden, CO 80401, USA*

[7]*Department of Chemistry and Institute of Materials Science and Engineering, Washington University in St. Louis, St. Louis, MO 63130, USA*

(Dated: May 25, 2026)

Abstract

Oxygen vacancy formation energies govern the performance of metal oxides across energy conversion, catalysis, and electronics, yet predicting them accurately without density functional theory calculations for novel chemical systems remains challenging. Here, we show that frozen 256-dimensional embeddings extracted from a pretrained MACE universal machine-learning interatomic potential encode information sufficient to predict vacancy formation energies obtained from density functional theory without requiring supercell construction, vacancy creation, or geometry optimization. A lightweight multilayer perceptron trained on these embeddings (MACE-dGNN) achieves an element-wise cross-validation mean absolute error of 0.38 eV, halving the error of a baseline graph neural network trained from scratch (0.72 eV) and offering a modest advantage over direct MACE relaxation calculations (0.51 eV). Because the embedding approach decouples representation from the prediction task and generalizes well with limited data, it will (1) extend naturally to fine-tuning and prediction of properties (e.g., charged defects) that are inaccessible to generic energy- and force-output interatomic potentials and (2) be trainable on necessarily small datasets, potentially enabling the use of more expensive but accurate tools for generating training data (e.g., hybrid functional calculations) for which training or fine-tuning of interatomic potentials will be difficult. We demonstrate practical impact by revisiting a prior screening of thermochemical water-splitting materials, where improved generalization alters ∼45 % of candidate classifications.

I. INTRODUCTION

Oxygen vacancies are among the most extensively studied point defects in metal oxides, governing performance across application areas in energy conversion, catalysis, and electronics. The energy required to form an oxygen vacancy, by removing a lattice oxygen atom and accommodating the resulting structural and electronic reorganization, controls ionic conductivity in solid oxide fuel cells and electrolyzers, [1–4] catalytic activity through the Mars–van Krevelen mechanism, [5–8] resistive switching behavior in memristors and neuromorphic devices, [9–11] redox thermodynamics in thermochemical water-splitting cycles, [12] and the properties of optically active defect centers relevant to quantum information sci-

* mwitman@sandia.gov

† wexler@wustl.edu

ence. [13–15] The optimal vacancy formation energy differs substantially across applications, ranging from $\approx 0.5-1.5\,\mathrm{eV}$ for ionic conductors to $\approx 2-4\,\mathrm{eV}$ for thermochemical redox materials [12, 16–18], yet the chemical space of candidate metal oxides is vast. Navigating this design space requires predictive models that can rapidly screen tens of thousands of known and hypothetical oxides in databases such as the Materials Project. [19, 20]

While density functional theory (DFT) calculations provide reliable vacancy formation energies, the cost of relaxing defective supercells, which generally contain hundreds of atoms, limits their throughput. Surrogate models that predict defect energetics from host-structure information alone are therefore valuable for high-throughput screening. A critical requirement for such models is out-of-sample generalization: because materials discovery tasks often target compositions, space groups, chemical systems, etc., absent from the training data [21], stricter performance metrics can be defined than just accuracy on randomly held-out structures, e.g., accuracy in element-wise cross-validation (CV) test splits in which all structures containing a given cation are withheld from training. In prior work, we introduced a defect graph neural network (dGNN) that adapts the crystal graph convolutional neural network (CGCNN) architecture [22] to predict site-resolved vacancy formation energies from the host crystal graph. [23] Trained on ∼1,900 DFT-computed neutral vacancy energies in metal oxides (of which ∼1,100 are oxygen vacancies and the rest cation vacancies), the dGNN achieves a structure-wise CV mean absolute error (MAE) of $\approx 0.38\,\mathrm{eV}$ for oxygen vacancies and has enabled experimental discovery of thermochemical water-splitting materials with ∼80 % true-positive rates. [12] However, the element-wise MAE roughly doubles to $\approx 0.75\,\mathrm{eV}$, [12] reflecting the fundamental difficulty of extrapolating to underrepresented or novel chemistries from a small, specialized training set.

Universal machine-learning interatomic potentials (uMLIPs) offer a potential solution. Foundational models such as MACE, [24–26] CHGNet, [27] and M3GNet [28] are trained on hundreds of thousands to millions of structures spanning much of the periodic table, encoding chemical knowledge that is unavailable to specialized models trained on small, curated datasets. They provide two routes to potentially improve element-wise generalization. First, *if* the target property prediction task relies only on structure inputs (atomic positions and lattice vectors) and force/energy outputs, such as the neutral vacancy formation energies studied herein, uMLIPs can be applied in zero-shot with no additional training required; this approach provides high-quality predictions for data points that are chemically and struc-

turally in-distribution relative to the uMLIP's large training data, but would otherwise be out-of-distribution relative to small, domain-specific curated datasets needed for bespoke model training. The second alternative is to ask whether the latent space (specifically, crystal site embeddings) learned by uMLIPs across diverse structure and chemistry space can be repurposed by training low-complexity downstream models on these embeddings, and whether this could yield better element-wise out-of-distribution generalization than direct uMLIP relaxations or bespoke, trained-from-scratch models (dGNN).

Indeed, a growing body of work demonstrates that the embeddings learned by these potentials can be extracted and repurposed for downstream property prediction. The DPA-1 model enables efficient fine-tuning for defect and surface energies when pretrained on large datasets. [29] The HackNIP pipeline extracts embeddings from pretrained potentials for use in shallow machine-learning models, [30] while the "franken" framework adapts MACE-MP0 atomic descriptors via kernel methods to predict surface and defect energies. [31] Δ-learning approaches on internal uMLIP representations can further correct systematic errors for challenging systems. [32] These studies establish that foundational uMLIP embeddings encode chemically meaningful, transferable information; however, their effectiveness depends on the similarity between source and target domains, and systematic errors can arise for high-energy configurations underrepresented in bulk-dominated training sets. [33, 34] Whether foundational uMLIP embeddings specifically improve predictions of oxygen vacancy formation energies, and whether such improvement manifests in the element-wise generalization regime, has not been systematically investigated.

Here, we address this question by extracting 256-dimensional embedding vectors from a pretrained MACE uMLIP [24–26] for crystallographic sites in host structures and using them as input features to predict DFT-calculated vacancy formation energies. We compare a hierarchy of structure-property, MACE-derived models, ranging from linear regression on scalar site energies (MACE-E_i), through linear regression on full embedding vectors (MACE-LR), to a multilayer perceptron (MLP) on embeddings (MACE-dGNN), against the original dGNN [23], which learns representations from scratch, and against a workflow using direct MACE relaxations to compute vacancy formation energies (MACE-rlx). We evaluate all structure-property models using nested CV with both structure-wise and element-wise data splits, focusing on the latter as the most stringent test of generalization for materials discovery. Without requiring supercell construction, vacancy creation, or geometry optimization,

MACE-dGNN achieves an element-wise MAE of 0.38 eV, modestly outperforming the direct MACE relaxation workflow (0.51 eV). This establishes that pretrained uMLIP embeddings encode sufficient information about local defect environments to support structure-property models, even for compositions or chemical systems absent from the surrogate model's training data. Simultaneously, the ∼50 % reduction relative to the baseline dGNN (0.72 eV) demonstrates the practical value of pretrained representations for out-of-sample generalization from small, chemically nonuniform datasets.

Because the embedding approach decouples the learned representation from the prediction target, the approach should extend to properties where direct uMLIP energy differences are inapplicable: for instance, charged-defect transition levels that depend on the Fermi level and require finite-size corrections [35–37], vacancy-migration barriers that require nudged elastic band calculations on defective supercells [38, 39], and properties requiring accuracy exceeding that of standard density functionals based on the generalized gradient approximation, e.g., the functional of Perdew, Burke, and Ernzerhof (PBE) [40]. Kiyohara et al. [35] have already demonstrated charged-defect property predictions via a dGNN surrogate model trained from scratch. Finally, we revisit the thermochemical water-splitting screening of Ref. [12] with MACE-dGNN to assess the practical impact of improved generalization on materials discovery outcomes. While demonstrated here for neutral vacancy formation energies, the methodology should be general and extensible to other defect types, material classes, and oxide properties.

II. RESULTS

A. Vacancy formation energy training data

This paper addresses the surrogate modeling task of computing the neutral vacancy formation energy at a given atomic site in a host crystal structure. The formation enthalpy $\Delta H_{V_{\mathrm{X},i}}$ for an uncharged vacancy of species X (i.e., cation or oxygen) is defined from DFT total energies as follows: [37]

$$\Delta H^{\mathrm{DFT}}_{V_{\mathrm{X},i}} = E_{V_{\mathrm{X},i}} - E_{\mathrm{bulk}} + \mu_{\mathrm{X}}. \tag{1}$$

Here, $E_{V_{\mathrm{X},i}}$ is the DFT-computed total energy of a supercell containing the neutral vacancy $V_{\mathrm{X},i}$; E_{bulk} is the total energy of the pristine, defect-free supercell; and $\mu_{\mathrm{X}} = \mu^{\mathrm{ref}}_{\mathrm{X}} + \Delta\mu_{\mathrm{X}}$ is

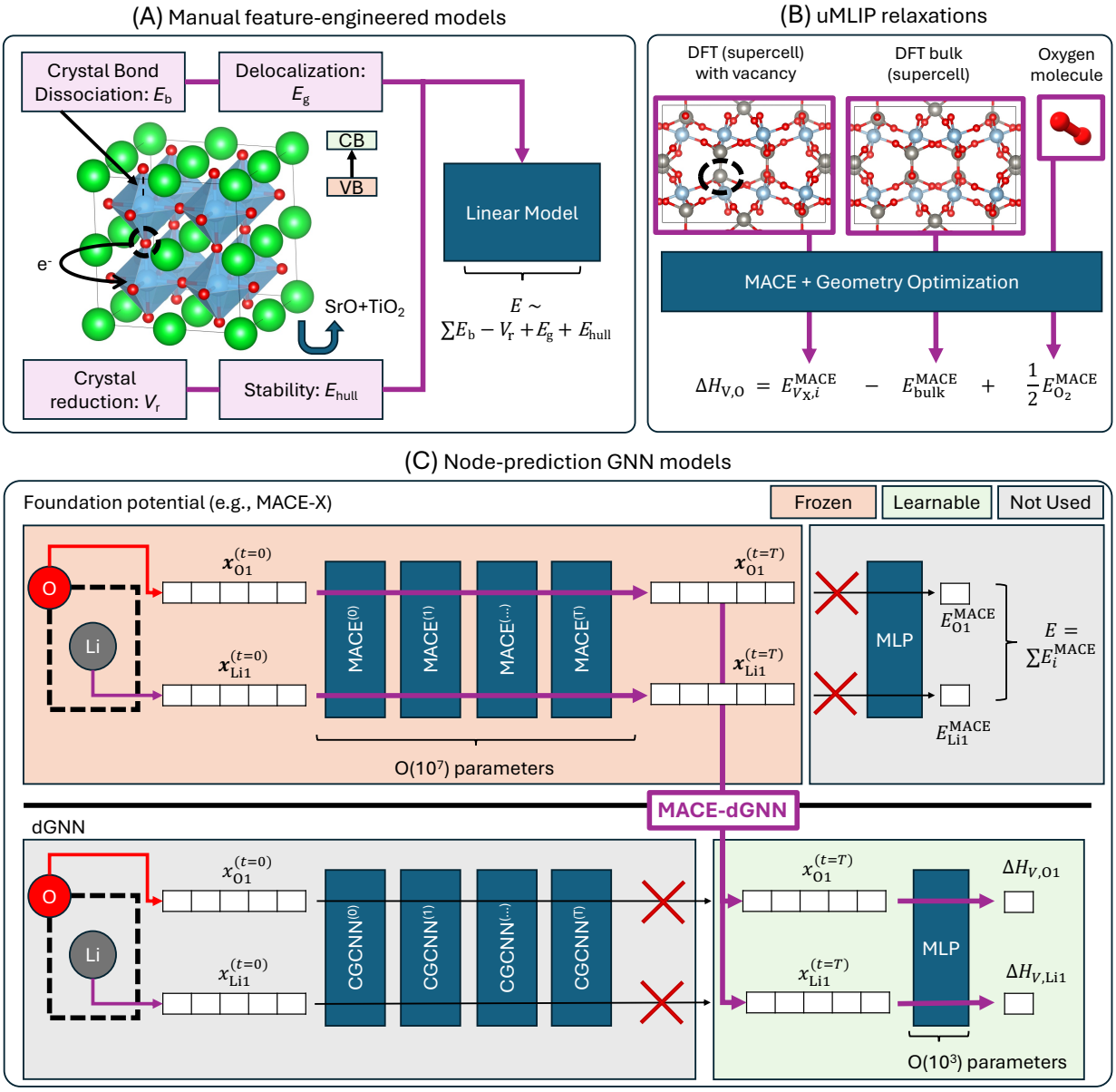

FIG. 1. Schematic examples of surrogate model alternatives to DFT for predicting point vacancy formation energies in crystalline materials. (a) Linear models using hand-selected features (E_g = band gap, E_{hull} = canonical energy above the hull, E_b = crystal bond dissociation energies, V_r = crystal reduction potential); (b) Direct use of a uMLIP (such as the MACE family of models) to perform structural relaxations needed to evaluate Equation (1). (c) A dGNN-style structure-property model that uses host structures as input (no relaxations) and uses the per-site representations produced by T frozen MACE message-passing steps $\text{MACE}^{(t=0\ldots T)}$ to generate per-site embeddings that serve as fixed inputs to a simple MLP trained for direct ΔH_{V_X} predictions.

the elemental chemical potential of X. The reference $\mu_{\mathrm{X}}^{\mathrm{ref}}$ is the total energy of the elemental phase of the removed element (e.g., molecular O_2 for oxygen). The additional term $\Delta\mu_{\mathrm{X}}$ represents conditions in the specific experimental environment and must be non-positive for the host crystal structure to be thermodynamically stable relative to the elemental phase. For gas phases, it can be expressed as function of the temperature and the partial pressure, e.g., $\Delta\mu_{\mathrm{O}} = \Delta\mu_{\mathrm{O}}(T, p_{\mathrm{O}_2})$ (see, e.g., Ref. [18]). Additional bounds on the $\Delta\mu_{\mathrm{X}}$ are placed by the thermodynamic stability of other limiting phases in the elemental phase space. Because $\Delta\mu_{\mathrm{X}}$ is a free variable not directly connected to defect calculations, for the purposes of screening, we consider the case $\Delta\mu_{\mathrm{X}} = 0$ ($\mu_{\mathrm{X}} = \mu_{\mathrm{X}}^{\mathrm{ref}}$) in the following discussion except where noted.

Our $\Delta H_{V_{\mathrm{X},i}}$ training database is identical to that of [12], which merged data from [23, 41] with some additional calculations detailed therein. Specific DFT details on pseudopotentials, exchange–correlation functionals, supercell sizes, etc., can be found in Refs. [12, 23]. The database, also provided in supplementary files, contains $\Delta H_{V_{\mathrm{X},i}}$ for ∼1,900 unique crystallographic sites (of which ∼1,100 are oxygen) spanning ∼250 unique compounds. However, the coverage of elemental space in this database is not uniform, with several cations represented only by a single binary oxide structure. This imbalance, and its implications for improved model generalization, is quantified in Section II B 2 and discussed as a limitation in Section III.

B. Surrogate modeling approaches

Our first goal is to develop surrogate approaches to computing Equation (1) and compare them to the baseline dGNN [23]. The approaches investigated in this section are listed in Table I in chronological order, along with their model type and required input.

1. MACE-E_i: host site energies provide physical insights into oxygen vacancy thermodynamics

We postulate that the per-site scalar energies output by a uMLIP such as MACE (i.e., the values produced by the final linear projection applied to each site's pre-readout embedding, hereafter the "readout"; e.g., $E_{\mathrm{bulk},i}^{\mathrm{MACE}}$ in Fig. 1c) for oxygen sites in the pristine bulk should correlate with oxygen vacancy formation energies, $\Delta H_{V_{\mathrm{O}}}^{\mathrm{DFT}}$. Here we choose the `mace-mh-0`

Model name	Brief description	Model input
dGNN	Baseline GNN model [12, 23]	DFT-relaxed host crystal structure
MACE-E_i	Single-variable linear regression model	MACE site energy
MACE-LR	256-dimensional linear regression model	MACE site embedding (after message passing)
MACE-dGNN	Two-layer (64×64) multilayer perceptron	MACE site embedding (after message passing)
MACE-sp	MACE-calculated single-point energy differences between host and vacancy-removed supercells	DFT-relaxed host crystal structure
MACE-rlx	Full MACE relaxation workflow to compute Equation (1)	MACE-relaxed host and defect supercells

TABLE I. Surrogate models used to compute vacancy formation energies in this work, listed in the order they are investigated in Section II.

model with the head layer trained on the Open Materials 2024 (OMAT) dataset [42] under the PBE/PBE+U exchange–correlation functional (hereafter the OMAT PBE head), to provide inference predictions most aligned with the level of DFT theory used to create our dataset; performance of additional MACE models is investigated in the Supplementary Information (SI). E_i^{MACE} are the per-site atomic energies that, when summed, yield the total predicted energy for a given structure: $E_{\mathrm{tot}}^{\mathrm{MACE}} = \sum_i E_i^{\mathrm{MACE}} = \sum_i \left(e_i^{\mathrm{MACE}} + \epsilon_{\mathrm{X}_i}\right)$. Thus E_i^{MACE} is the sum of e_i^{MACE}, the learned node or "interaction" energy for site i, and ϵ_{X_i}, the constant reference energy that depends only on the site's elemental identity, X_i. Figure 2a shows a positive correlation between $\Delta H_{V_{\mathrm{O}}}^{\mathrm{DFT}}$ and $-E_{\mathrm{bulk},i}^{\mathrm{MACE}} + 1/2 E_{\mathrm{O}_2}^{\mathrm{MACE}}$. A single-variable linear regression model, which we denote MACE-E_i, follows directly (though without sufficient accuracy):

$$\Delta H_{V_{\mathrm{O},i}}^{\text{MACE-}E_i} = 1.20\left(-E_{\mathrm{bulk},i}^{\mathrm{MACE}} + 1/2 E_{\mathrm{O}_2}^{\mathrm{MACE}}\right) + 0.48. \tag{2}$$

Preceding E_i^{MACE} in the uMLIP predictions, however, is the post-message-passing, pre-

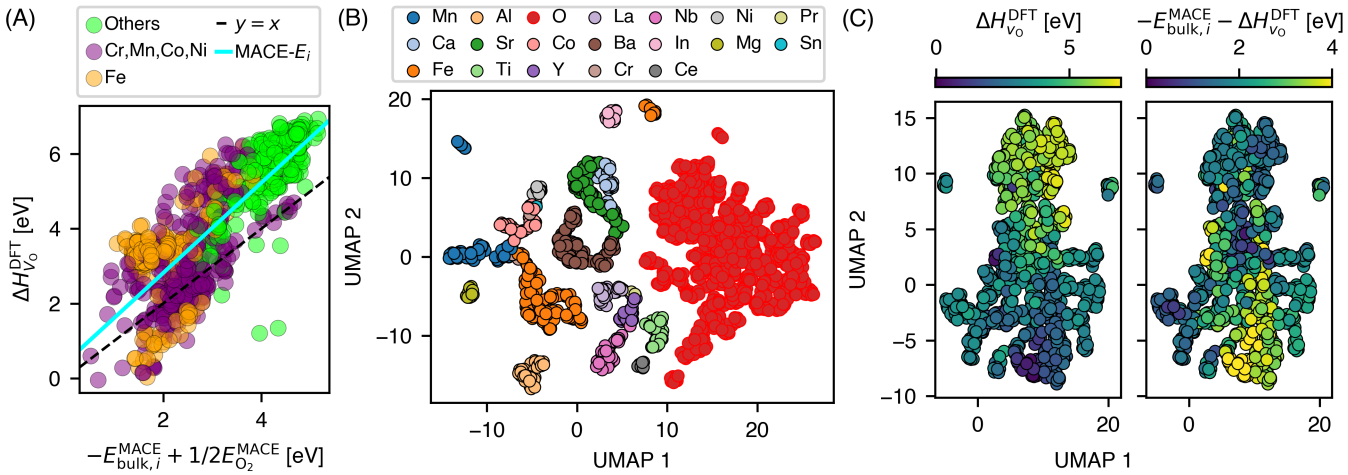

FIG. 2. (a) $\Delta H_{V_O}^{\mathrm{DFT}}$ vs. $-E_{\mathrm{bulk},i}^{\mathrm{MACE}} + 1/2E_{\mathrm{O}_2}^{\mathrm{MACE}}$, color-coded by whether their host structure contains Fe (orange), several other first-row transition metals (purple), or none of the above (green). Black line represents $y = x$ while the cyan line represents the line of best fit, Equation (2). (b) UMAP dimensionality reduction of $\boldsymbol{x}_i^{\mathrm{MACE}}$ for all vacancy sites in the database, color-coded by the elemental site identity. (c) UMAP dimensionality reduction of $\boldsymbol{x}_i^{\mathrm{MACE}}$ for all oxygen sites in the database, color-coded by $\Delta H_{V_O}^{\mathrm{DFT}}$ or $-E_i^{\mathrm{MACE}} - \Delta H_{V_O}^{\mathrm{DFT}}$.

readout feature vector $\boldsymbol{x}_i^{\mathrm{MACE}}$, which encodes information about each site's local environment in a low-dimensional, non-interpretable feature space. Performing a dimensionality reduction on $\boldsymbol{x}_i^{\mathrm{MACE}}$ using Uniform Manifold Approximation and Projection (UMAP) shows the following. The 2-component UMAP reduction in Figure 2b shows that cation sites tend to form relatively compact, non-overlapping clusters, while oxygen sites are non-overlapping with cation sites but are much more diffuse in the UMAP embedding space. Notably, Fe and Mn yield the most diffuse clusters among first row transition elements, as expected based on their diversity of possible oxidation states.

Figure 2c further color-codes a 2-component UMAP reduction of just oxygen sites by $\Delta H_{V_O}^{\mathrm{DFT}}$, revealing the local grouping of low and high vacancy formation energy sites. Thus, because $-E_{\mathrm{bulk},i}^{\mathrm{MACE}}$ is less correlated to $\Delta H_{V_O}^{\mathrm{DFT}}$ for small values of $\Delta H_{V_O}^{\mathrm{DFT}}$, color-coding each site by $-E_{\mathrm{bulk},i}^{\mathrm{MACE}} - \Delta H_{V_O}^{\mathrm{DFT}}$ also shows local clustering. These results indicate that $\boldsymbol{x}_i^{\mathrm{MACE}}$ can serve as a descriptor for structure-property defect models, as examined in Section II B 2. The same embedding space can also support training-database expansion: candidate defect sites projected into the MACE embedding space that occupy sparse or unrepresented regions can

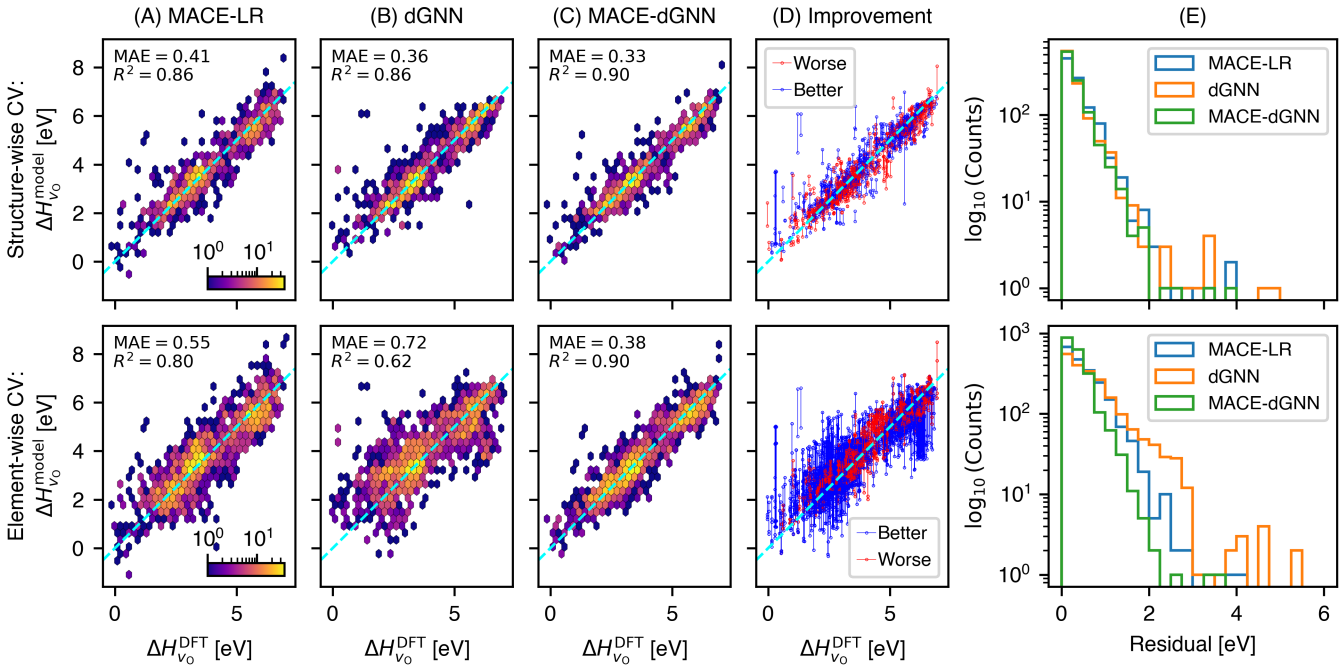

FIG. 3. Concatenation of all outer test set predictions from nested K-fold CV for the three model types considered herein: (a) MACE-LR, (b) dGNN, and (c) MACE-dGNN, for both structure-wise (top) and element-wise (bottom) hold-out strategies. Test set performance is reported only for the subset of predictions corresponding to oxygen vacancies (the full-dataset counterpart, including cation vacancies, is shown in the SI); expectation values are computed over the outer test sets. (d) MACE-dGNN's improved (or worsened) individual test set predictions relative to dGNN. (e) Histogram of residuals for all test set predictions.

be prioritized for DFT calculation, analogous to strategies that use pretrained embeddings to reduce redundant sampling in interatomic potential training [43–46].

2. MACE-LR and MACE-dGNN: Transfer learning on embeddings improves baseline predictions

We seek a structure-property surrogate model for Equation (1) that does not require manual feature engineering (in contrast to the linear or tree-based models of Refs. [36, 41, 47], schematically shown in Fig. 1a, which are generally more interpretable but less flexible). Our baseline deep learning model for structure-property defect prediction is the dGNN proposed in Ref. [23] and schematically shown in Figure 1c. Trained from scratch using CGCNN

convolutions followed by a readout, node-wise MLP, the dGNN predicts neutral vacancy (V_{X}) formation energies as

$$\Delta H_{V_{\mathrm{X},i}} = f_{\mathrm{dGNN}}(\mathcal{C}_h, i; \theta), \tag{3}$$

where $\mathcal{C}_h$ denotes the perfect (relaxed) host crystal structure, i the index of the crystallographic site hosting the vacancy, X the elemental identity of that site, and θ the learned model weights. The exact model architecture and training hyperparameters are provided in the SI. This architecture has been extended to more complex defect modeling tasks, including vacancy migration energies [38, 39] and charged vacancy formation energies [35].

Rather than training dGNN to learn local environments from scratch through its own message-passing layers, we now investigate whether the information encoded in $\boldsymbol{x}_i^{\mathrm{MACE}}$ yields improved predictions for $\Delta H_{V_{\mathrm{X},i}}^{\mathrm{DFT}}$. Here we use $\boldsymbol{x}_i^{\mathrm{MACE}} \in \mathbb{R}^{256}$ from the `mace-omat-0-small` model (instead of the `mace-mh-0` model) as our input feature vector, given the small size of the dataset available for downstream training. For clarity, we summarize only test-set performance on oxygen vacancies in the main manuscript, but all downstream models were trained on the entire dataset (cation and oxygen vacancies), and CV performance across all vacancy types is summarized in the SI.

We first investigate the utility of $\boldsymbol{x}_i^{\mathrm{MACE}}$ as the input for a linear regression (LR) model, denoted MACE-LR,

$$\Delta H_{V_{\mathrm{X},i}} = \left(\boldsymbol{x}_i^{\mathrm{MACE}}\right)^\top \boldsymbol{\beta}, \tag{4}$$

where $\boldsymbol{\beta}$ are the learned linear model coefficients.

For a more expressive model that is more conceptually similar to dGNN, we use a minimal-complexity MLP for node-wise, readout predictions, denoted here as MACE-dGNN and schematically shown in Figure 1c,

$$\Delta H_{V_{\mathrm{X},i}} = \sigma_L(\boldsymbol{W}_L \sigma_{L-1} \ldots (\sigma_1(\boldsymbol{W}_1 \boldsymbol{x}_i^{\mathrm{MACE}} + \boldsymbol{b}_1) \cdots + \boldsymbol{b}_{L-1}) + \boldsymbol{b}_L). \tag{5}$$

Here, $\boldsymbol{W}_l$ are learnable weight matrices, $\boldsymbol{b}$ are bias vectors, and σ_l are activation functions. More specific architecture details are given in the SI.

For each of these three models, Figure 3a–c shows the outer test set predictions from (K, L)-fold nested CV for two different criteria for allocating train/test splits (see SI for details). The top row represents $K = 10$ "Structure-wise" CV, where all defects from a given host structure are randomly assigned to the test set. The bottom row represents $K = 18$

leave-one-out "Element-wise" CV, where all defects from any host structure that contains the target test cation are assigned to the test set (with the exception of any binary oxide composition, which is always assigned to the train set). For "Structure-wise" CV, where train and test distributions much more closely overlap, the performance gain from MACE-dGNN is minimal in comparison to the baseline dGNN, although a handful of outlying errors in dGNN are improved by MACE-dGNN, as shown in Figure 3d, with the distribution of residuals specifically quantified in Figure 3e. For "Element-wise" CV, the improvement is substantially larger, with a corresponding reduction in the element-wise generalization gap.

3. MACE-sp: single-point energy differences yield strong zero-shot performance

We next investigate the predictive capability of MACE when used only as a many-body potential energy calculator to compute a single-point vacancy formation energy (MACE-sp),

$$\Delta H_{V_{\mathrm{O},i}}^{\text{MACE-sp}} = E_{V_{\mathrm{O}},i}^{\text{MACE-sp}} - E_{\text{bulk}}^{\text{MACE-sp}} + \mu_{\mathrm{O}}^{\text{MACE}}. \tag{6}$$

Here, $E_{\text{bulk}}^{\text{MACE-sp}}$ indicates a MACE single-point total energy calculation on the *DFT-relaxed* bulk supercell (same input as dGNN, MACE-LR, and MACE-dGNN), $E_{V_{\mathrm{O}},i}^{\text{MACE-sp}}$ indicates a MACE single-point total energy calculation on the defected supercell after removing one atom to create a vacancy but *without* relaxation, and $\mu_{\mathrm{O}}^{\text{MACE}} \equiv 1/2 E_{\mathrm{O}_2}^{\text{MACE}}$.

The total energies in Equation (6) can be expanded into their respective per-site energy summations,

$$\Delta H_{V_{\mathrm{O},i}}^{\text{MACE-sp}} = \sum_{j \neq i} E_{V_{\mathrm{O}},j}^{\text{MACE}} - \left(\sum_{j \neq i} E_{\text{bulk},j}^{\text{MACE}} + E_{\text{bulk},i}^{\text{MACE}} \right) + 1/2 E_{\mathrm{O}_2}^{\text{MACE}}, \tag{7}$$

and rearranged to yield

$$\Delta H_{V_{\mathrm{O},i}}^{\text{MACE-sp}} = \Delta_j^{\text{MACE}} - E_{\text{bulk},i}^{\text{MACE}} + 1/2 E_{\mathrm{O}_2}^{\text{MACE}}. \tag{8}$$

In this expression,

$$\Delta_j^{\text{MACE}} = \sum_{j \neq i} \delta_j^{\text{MACE}} = \sum_{j \neq i} \left(E_{V_{\mathrm{O}},j}^{\text{MACE}} - E_{\text{bulk},j}^{\text{MACE}} \right) \tag{9}$$

is the neighbor response term, i.e., the summation across all non-vacancy sites' energy differences, δ_j^{MACE}, between the pristine and defected supercell.

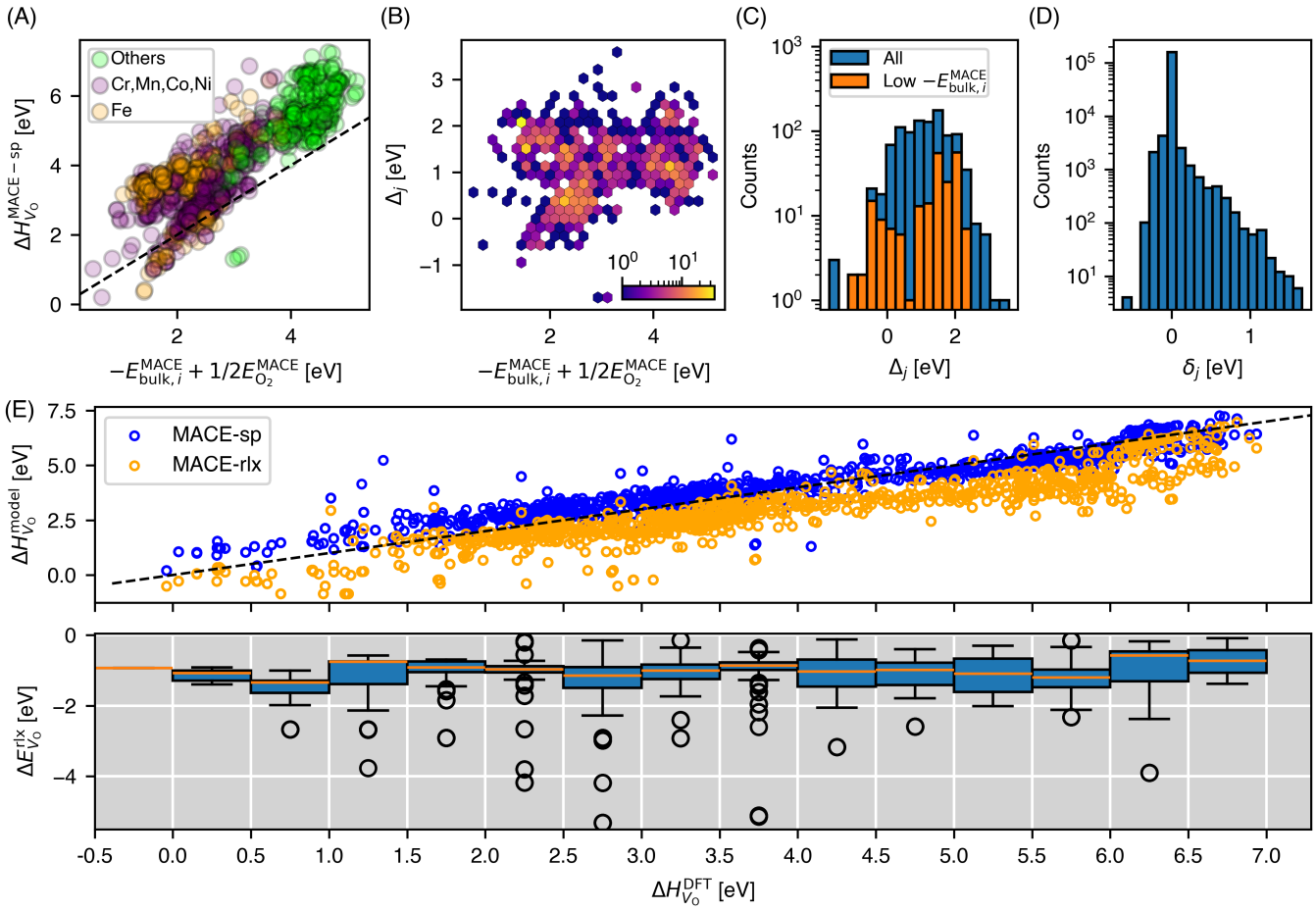

FIG. 4. (a) $\Delta H^{\mathrm{MACE\text{-}sp}}_{V_{\mathrm{O},i}}$ vs. $-E^{\mathrm{MACE}}_{\mathrm{bulk},i} + 1/2E^{\mathrm{MACE}}_{\mathrm{O}_2}$, color-coded by elements present in the compound. Dashed black line represents $y = x$. (b) Δ_j vs. $-E^{\mathrm{MACE}}_{\mathrm{bulk},i} + 1/2E^{\mathrm{MACE}}_{\mathrm{O}_2}$, color-coded by data density (counts). (c) Distribution of Δ_j across all MACE-sp vacancy site calculations, as well as the subset where $-E^{\mathrm{MACE}}_{\mathrm{bulk},i} + 1/2E^{\mathrm{MACE}}_{\mathrm{O}_2} < 2\,\mathrm{eV}$. (d) Distribution of δ_j across all non-vacancy sites for all MACE-sp vacancy calculations. (e) Unshifted MACE-sp and MACE-rlx predictions are plotted vs. $\Delta H^{\mathrm{DFT}}_{V_\mathrm{O}}$, while the boxplots beneath quantify the distribution of defect supercell relaxation energies across all defects within each $\Delta H^{\mathrm{DFT}}_{V_\mathrm{O}}$ bin.

Figure 4a shows the MACE-sp model predictions vs. $-E^{\mathrm{MACE}}_{\mathrm{bulk},i} + 1/2E^{\mathrm{MACE}}_{\mathrm{O}_2}$. The x-axis can also be interpreted as the difference between the interaction energy of an oxygen atom in its reference state (O_2 molecule) and in the pristine host. With predictions color-coded in the same way as in Figure 2a, the qualitative similarity between the two plots indicates that MACE-sp serves as a strong surrogate model for $\Delta H^{\mathrm{DFT}}_{V_\mathrm{O}}$, as quantified in Section II C.

The deviation of $\Delta H^{\mathrm{MACE\text{-}sp}}_{V_{\mathrm{O},i}}$ from $y = x$ in Figure 4a corresponds to the Δ_j term. While

Δ_j is approximately symmetrically distributed across *all* defect sites (Fig. 4c), the behavior is more complex for vacancies with low formation energies, $\Delta H^{\text{MACE-sp}}_{V_{\text{O},i}} \lesssim 4\,\text{eV}$. Here, the distribution of Δ_j broadens and becomes bimodal as $-E^{\text{MACE}}_{\text{bulk},i} + 1/2E^{\text{MACE}}_{\text{O}_2}$ decreases below 2.5 eV. This cluster of low Δ_j vacancy sites corresponds exclusively to materials containing first-row transition metals (Cr, Mn, Fe, Co, Ni), which exhibit variable oxidation states. More broadly, this observation suggests that the MACE site-energy decomposition is physically meaningful. Although artificial with respect to DFT, this decomposition of per-site bulk energies $E^{\text{MACE}}_{\text{bulk},i}$ is known to meaningfully capture the local chemical environment of sites in the pristine host, precisely because MACE accurately captures total energies (as validated against DFT-derived formation energies elsewhere [26]). Meanwhile, the neighbor response term Δ_j is only weakly correlated with $E^{\text{MACE}}_{\text{bulk},i}$ and captures the many-body energetic response to removing an oxygen atom while the surrounding site energies readjust at fixed geometry. The sum $-E^{\text{MACE}}_{\text{bulk},i} + \Delta_j$ (Equation (7)) yields a strong surrogate model for $\Delta H^{\text{DFT}}_{V_{\text{O}}}$, demonstrating MACE's ability to capture these many-body interactions in the host. For first-row transition-metal hosts, partially filled d shells permit multiple low-lying electronic configurations on the cation neighbors, so the multi-neighbor response to vacancy creation is large and varies sharply with the local environment, producing the bimodal Δ_j distribution observed at low $-E^{\text{MACE}}_{\text{bulk},i} + 1/2E^{\text{MACE}}_{\text{O}_2}$. Finally, Figure 4d shows that, as expected, the locality of point vacancy interactions yields $\delta_j \approx 0$ for most supercell sites, and that the δ_j distribution is highly skewed with exponentially decreasing probability of large δ_j values for neighboring sites to the vacancy.

4. MACE-rlx: direct relaxations yield strong zero-shot performance

Figure 4e shows that MACE-sp serves as a strong surrogate. However, this model assumes that a DFT-relaxed crystal structure is available as input. For high-throughput defect modeling coupled to hypothetical structure generation workflows, we investigate MACE's capability to reproduce the entire vacancy calculation workflow analogous to DFT in Equation (1). We compute the MACE-rlx model,

$$\Delta H^{\text{MACE-rlx}}_{V_{\text{O},i}} = E^{\text{MACE-rlx}}_{V_{\text{O}},i} - E^{\text{MACE-rlx}}_{\text{bulk}} + \mu^{\text{MACE}}_{\text{O}}, \tag{10}$$

where $E_{\text{bulk}}^{\text{MACE-rlx}}$ is the total energy of the pristine supercell after full MACE relaxation of cell lattice parameters and internal coordinates, and $E_{V_{\text{O}},i}^{\text{MACE-rlx}}$ is the total energy of the defected supercell after MACE internal coordinate-only relaxation.

Figure 4e's parity plot also shows the raw MACE-rlx predictions, which, like MACE-sp, correlate with $\Delta H_{V_{\text{O}}}^{\text{DFT}}$ but systematically underestimate it. This analysis omits two outlier oxygen vacancy sites (see SI) with anomalously low predicted formation energies that indicate a breakdown of the MACE relaxation of the defected supercell. Because direct MACE predictions use a different reference state and correction scheme than our underlying DFT data, quantitative comparison of MACE-sp and MACE-rlx to the other surrogate models is performed after applying a constant shift, $\mu_{\text{O}}^{\text{shift}}$, that minimizes the MAE of MACE relative to our DFT data (see Sec. V D).

We denote the shifted model predictions as MACE-sp-μ and MACE-rlx-μ, with optimal shifts of $-0.12\,\text{eV}$ and $0.91\,\text{eV}$, respectively. The optimal shift for MACE-rlx, however, is likely larger than can be explained by reference-state discrepancies between the data on which the potential was trained and our own (e.g., the Materials Project's correction scheme vs. our fitted elemental reference energies, FERE [48, 49]). Other discrepancies in DFT settings (e.g., U values) between OMAT and our dataset could in principle introduce further non-systematic, chemistry-specific errors that reduce the performance of the zero-shot MACE models (MACE-rlx and MACE-sp) referenced to our dataset (see SI for additional discussion). However, the R^2 of MACE-rlx-μ *decreases* relative to MACE-sp-μ. This indicates that the MACE potential energy surface around the defect state is too shallow, with MACE relaxations descending to lower energies than the corresponding DFT relaxations. Figure 4e shows, as a function of bins in $\Delta H_{V_{\text{O}}}^{\text{DFT}}$, the boxplot distributions of the defect supercell relaxation energies, $\Delta E_{V_{\text{O}}}^{\text{rlx}}$ (outliers less than $-5.5\,\text{eV}$ not visualized). The medians of the $\Delta E_{V_{\text{O}}}^{\text{rlx}}$ distributions are approximately constant with $\Delta H_{V_{\text{O}}}^{\text{DFT}}$, indicating a systematic, roughly constant over-relaxation of the defect states. However, a substantial number of low-$\Delta E_{V_{\text{O}}}^{\text{rlx}}$ outliers drive the MACE-rlx underestimation of $\Delta H_{V_{\text{O}}}^{\text{DFT}}$ for several vacancy sites.

C. Improving accuracy and reducing bias for out-of-distribution predictions

Figure 5 provides a comparative performance analysis across all modeling approaches for "Structure-wise" and "Element-wise" CV. Box plots in Figure 5a–b visualize the distribution of MAE and R^2 across all outer K-fold test sets, $\{\text{MAE}\}_K$ and $\{R^2\}_K$, while stars visualize each metric computed across the concatenated test sets. The MACE-E_i, MACE-sp, and MACE-rlx models are zero-shot in the sense that none of their parameters are trained on the DFT vacancy database; they generate vacancy formation energies directly from the pre-trained MACE potential and a fitted scalar shift. For these zero-shot models, the box plots correspond to performance assessments on the same individual $\{\text{MAE}\}_K$ and $\{R^2\}_K$ test splits, allowing explicit comparison on the difficult-to-predict (particularly "Element-wise") splits. Across the full dataset, dGNN maintains competitive performance in "Structure-wise" CV (with MACE-dGNN slightly superior), i.e., when inference samples are relatively in-distribution (*id*) with respect to the chemical coverage of the training data.

The behavior changes substantially for "Element-wise" CV. Even the simplest model, MACE-E_i (a two-parameter linear regression), achieves performance comparable to the dGNN baseline. All other models improve substantially over dGNN, with MACE-sp, MACE-rlx, and MACE-dGNN performing best. Little distinction can be made between the relative performance of the top three models, although each shows elevated error on several test sets.

Figure 5c,d combines an analysis of performance for *id* vs. out-of-distribution (*ood*) chemistry space with performance for *id* vs. *ood* target property space. The histogram shows the distribution of $\Delta H_{V_\text{O}}^{\text{DFT}}$ target values above the averaged absolute error (AE) and averaged signed error (SE) of test set predictions within each $\Delta H_{V_\text{O}}^{\text{DFT}}$ bin for structure-wise and element-wise CV. All models show some performance degradation on low $\Delta H_{V_\text{O}}^{\text{DFT}}$ vacancies, but this is particularly pronounced for dGNN and especially so in element-wise CV. In contrast, MACE-dGNN exhibits lower expected AEs and expected SEs closer to zero for low and high $\Delta H_{V_\text{O}}^{\text{DFT}}$ vacancies.

D. Transfer learning's impact on applications

We investigate whether the CV improvements of MACE-dGNN translate into differences in an application-specific task. In the context of identifying novel metal oxide candidates for

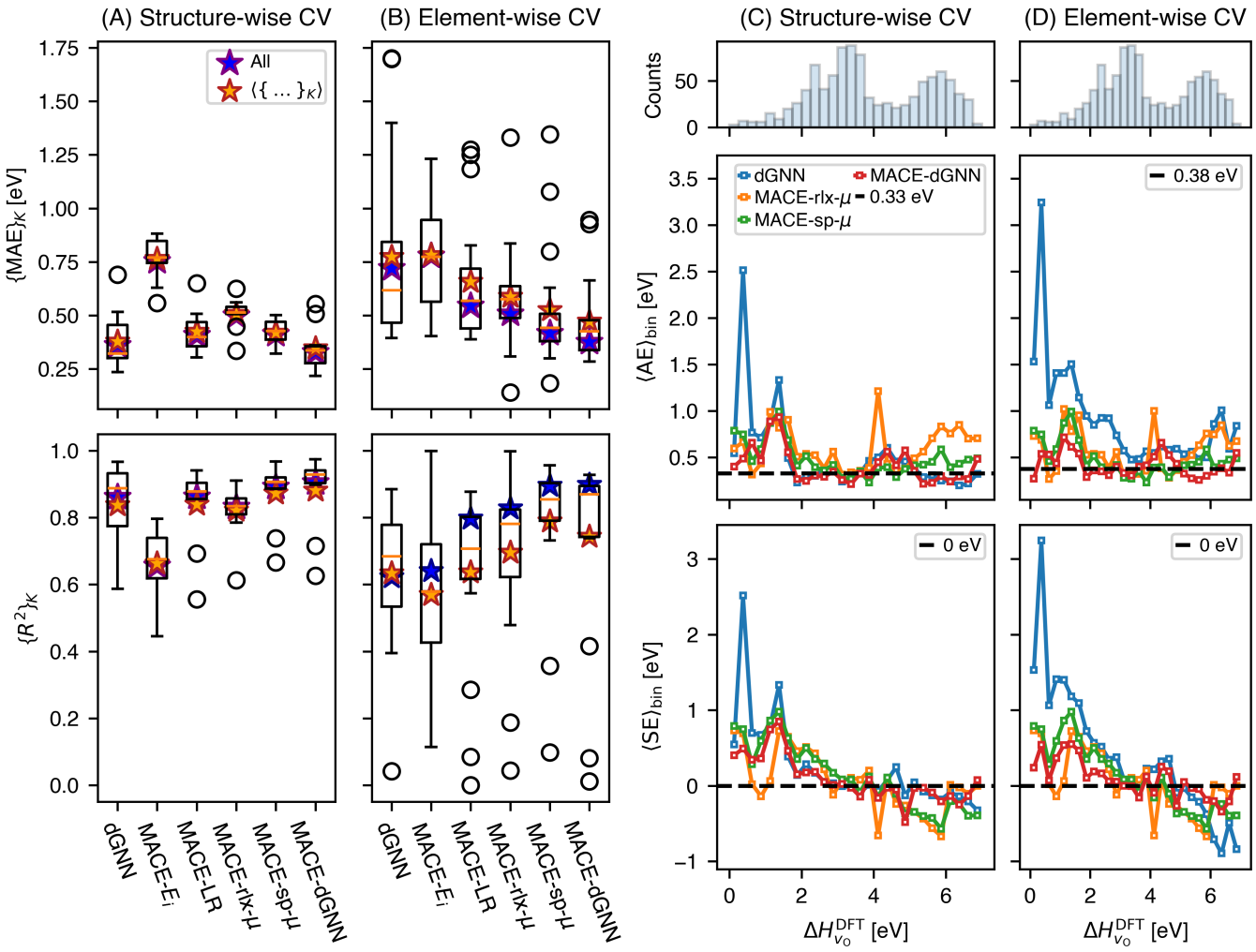

FIG. 5. Boxplot distributions of MAE and R^2 across all outer K-fold test sets, for (a) structure-wise CV and (b) element-wise CV. Blue stars represent an unweighted average of each metric across all test set predictions, while orange stars represent the ensemble average metric across K-folds. (c) Structure-wise CV and (d) Element-wise CV analysis of test set predictions showing histogram counts of the $\Delta H_{V_O}^{\mathrm{DFT}}$ target property above the averaged AE and SE of dGNN, MACE-rlx, and MACE-dGNN test set predictions within each bin. For AE plots, the dashed black line represents the MACE-dGNN global MAE, and for SE plots the black dashed line represents 0.

thermochemical water splitting (equivalently, thermochemical hydrogen production, TCH), we repeat the dGNN-based screening reported in Ref. [12] using the MACE-dGNN model. The strictness of down-selection criteria for screening metal oxides for TCH can vary, but here we adhere to two moderately strict down-selection criteria for identifying positive or screening hits.

	Structure-wise CV				Element-wise CV			
Model	MAE	$\langle\{\text{MAE}\}_K\rangle$	R^2	$\langle\{R^2\}_K\rangle$	MAE	$\langle\{\text{MAE}\}_K\rangle$	R^2	$\langle\{R^2\}_K\rangle$
dGNN	0.36	0.38	0.86	0.84	0.72	0.77	0.62	0.63
MACE-E_i	0.75	0.77	0.66	0.66	0.78	0.78	0.64	0.57
MACE-LR	0.41	0.42	0.86	0.84	0.55	0.66	0.80	0.64
MACE-rlx-μ	0.50	0.51	0.83	0.82	0.51	0.59	0.83	0.70
MACE-sp-μ	0.41	0.42	0.89	0.87	0.42	0.53	0.90	0.79
MACE-dGNN	0.33	0.35	0.90	0.88	0.38	0.47	0.90	0.74

TABLE II. Summary statistics for all models' performance, where values correspond to the blue (unweighted average across all test set predictions) and orange (ensemble average metrics across K-folds) stars in Figure 5a–b.

First, our surrogate model predicts the set of all oxygen vacancy formation energies in a given material, $\{\Delta H_{V_\text{O}}\}$, the minimum of which must satisfy

$$\{\Delta H_{V_\text{O}}\}_\text{min} \in [2.3, 4.0]\,\text{eV}, \tag{11}$$

for a positive hit. [23, 41, 50] Based on more stringent down-selection and TCH redox operating conditions, however, even narrower target ranges have been proposed, [16, 17] e.g., $[3.4, 3.9]\,\text{eV}$. Second, the host structure must be stable across the typical range of oxygen chemical potentials encountered during TCH redox, i.e., $[\Delta\mu_\text{O}^\text{TCH}] \approx [-2.5, -3.0]\,\text{eV}$ for favorable reduction and oxidation conditions.[51] As discussed when introducing Equation (1), these $\Delta\mu_\text{O}$ values refer to specific experimental conditions; in the case of oxygen, they correspond to temperatures and partial pressures of O_2 gas. For each host material, we define $\phi_h(\Delta\mu_\text{O})$ as the grand canonical energy above the hull as a function of the oxygen chemical potential. The range of $\Delta\mu_\text{O}$ over which $\phi_h(\Delta\mu_\text{O}) = 0$ (i.e., the host is on the convex hull) is denoted $[\Delta\mu_\text{O}^{\phi_h=0}]$. A host is TCH-stable when this range entirely spans the operating window $[\Delta\mu_\text{O}^\text{TCH}]$,

$$[\Delta\mu_\text{O}^\text{TCH}] \in [\Delta\mu_\text{O}^{\phi_h=0}]. \tag{12}$$

In practice, we require $[\Delta\mu_\text{O}^{\phi_h=0}]_\text{min} < -3.0\,\text{eV}$ for a positive hit. Note that $[\Delta\mu_\text{O}^{\phi_h=0}]$ can be computed from existing Materials Project data and its formation-energy correction schemes.

In Figure 6a–b, we show $[\Delta\mu_\text{O}^{\phi_h=0}]_\text{min}$ vs. $\{\Delta H_{V_\text{O}}\}_\text{min}$ for dGNN and MACE-dGNN screen-

ing predictions, which exhibit a negative correlation and an approximate lower bound of $-\{\Delta H_{V_O}\}_{\min} \approx [\Delta\mu_O^{\phi_h=0}]_{\min}$. This lower bound is physically meaningful: below it, the true chemical-potential-dependent formation energy of oxygen vacancies, $\{\Delta H_{V_O}\}_{\min} + [\Delta\mu_O^{\phi_h=0}]_{\min}$, would be negative, which is unphysical and suggests that the host structure is unstable. MACE-dGNN reduces the population of outlier materials with predicted $\{\Delta H_{V_O}\}_{\min} < -[\Delta\mu_O^{\phi_h=0}]_{\min}$, which we interpret as a reduction in erroneous screening predictions by dGNN.

Figure 6c plots the shift between dGNN- and MACE-dGNN-predicted $\{\Delta H_{V_O}\}_{\min}$, color-coded by sign and magnitude. For low (high) $\{\Delta H_{V_O}\}_{\min}$ and high (low) $[\Delta\mu_O^{\phi_h=0}]_{\min}$, dGNN systematically overpredicts (underpredicts) vacancy formation energies compared to MACE-dGNN. As shown in Figure 5, MACE-dGNN is more accurate (lower AE) and less biased (SE closer to zero) than dGNN across the entire target property range, particularly for *ood* materials whose compositions are underrepresented in the training data.

Figure 6d shows substantial differences between the two models when screening for negative and positive TCH hits. Of the 238 oxides for which Equations (11) and (12) are satisfied and thus identified as positives in the dGNN screening, 104 become negatives when using MACE-dGNN-predicted $\{\Delta H_{V_O}\}_{\min}$. Conversely, 110 oxides that were negatives in the dGNN screening become positives with the MACE-dGNN predictions. This is partly due to the high density of materials with $\{\Delta H_{V_O}\}_{\min}$ near the 4.0 eV upper limit, but many predictions vary by 1 eV or more between the two models. While the CV analysis in Figure 5 indicates that the MACE-dGNN predictions are more reliable, additional SI discussion of CV ensemble-based uncertainty further supports the improved performance of MACE-dGNN screening predictions.

Finally, from the screening of Ref. [12], we selected several compounds that contained cations with poor representation in the training data (Fig. 6e). We then computed their oxygen vacancy formation energies with DFT (see Methods for details), and compared them with both dGNN and MACE-dGNN predictions in Figure 6f. MACE-dGNN achieves a lower MAE than dGNN (0.43 eV vs. 0.54 eV, respectively) and correctly reclassifies one false negative as a true positive (Zn_2SnO_4) and two false positives as true negatives ($BaMoO_4$ and $FeWO_4$). This further validates the improved generalization performance of MACE-dGNN in *ood* prediction tasks.

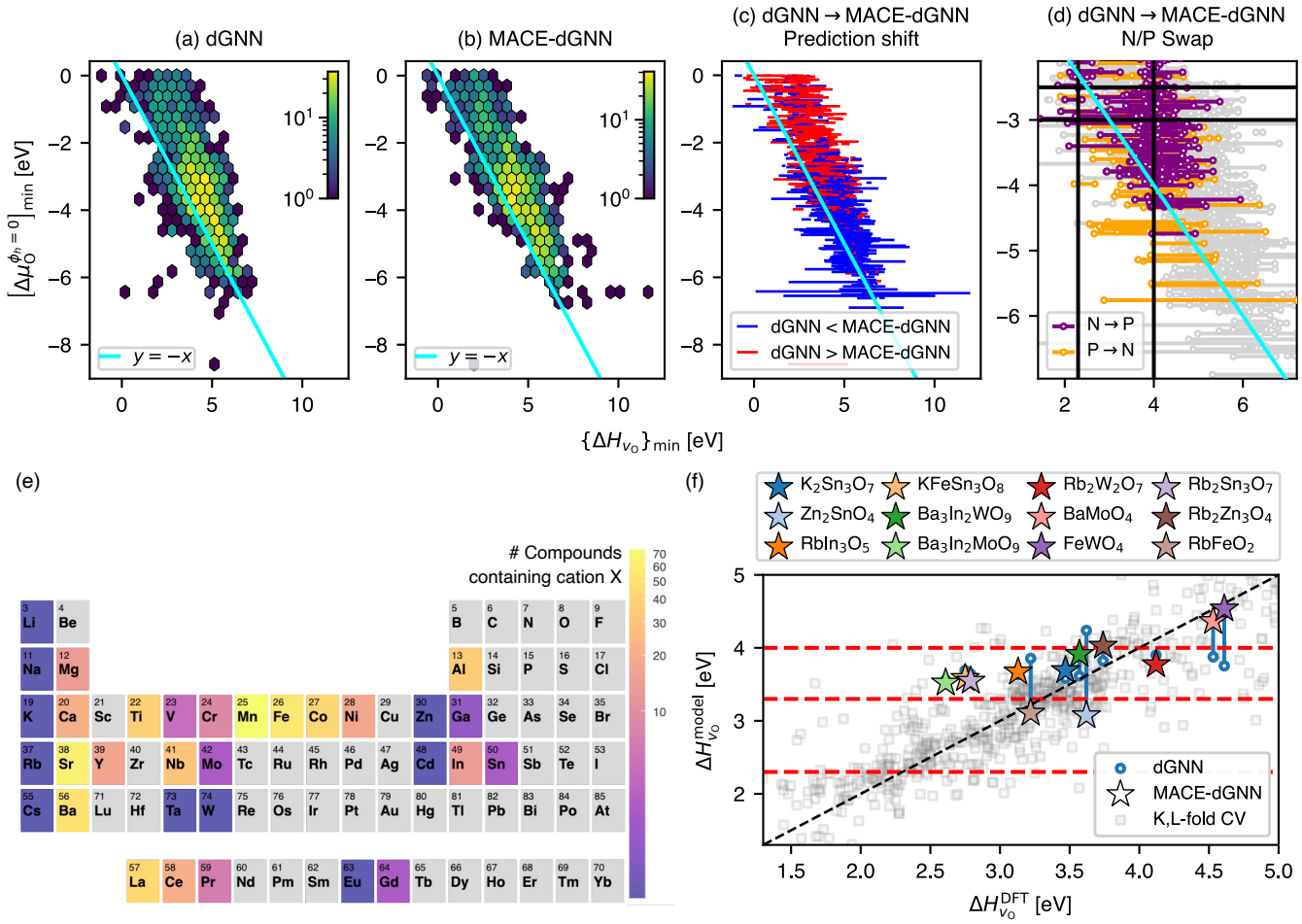

FIG. 6. Materials Project–computed $[\Delta\mu_O^{\phi_h=0}]_{min}$ for the oxide screening in Ref. [12] is plotted vs. (a) dGNN- and (b) MACE-dGNN-predicted $\{\Delta H_{V_O}\}_{min}$. The color bar corresponds to structure counts. For all subplots, the cyan line represents $y = -x$. (c) Horizontal lines connect dGNN and MACE-dGNN predictions of $\{\Delta H_{V_O}\}_{min}$ for each material, color-coded blue when the MACE-dGNN prediction is larger than the dGNN prediction and red when it is smaller. (d) Same as (c), except shifts are color-coded by whether MACE-dGNN converts a dGNN-predicted negative (N) to a positive (P) hit (purple), or vice versa (orange), for TCH down-selection; gray indicates that both models predict a negative TCH hit. The solid black lines mark the $\{\Delta H_{V_O}\}_{min}$ down-select window of $[2.3, 4.0]$ eV and the stability window $[\Delta\mu_O^{TCH}] = [-2.5, -3.0]$ eV. (e) Number of training compounds containing a given cation. (f) Original dGNN screening predictions (blue circles) for $\{\Delta H_{V_O}\}_{min}$, connected to revised MACE-dGNN predictions (colored stars), vs. the additional DFT validation calculations performed in this work. Gray squares show the structure-wise MACE-dGNN CV test predictions for comparison (i.e., data from Fig. 3c). Red dashed lines show the possible intervals for different TCH down-selection ranges, $[2.3, 4.0]$ eV and $[3.4, 3.9]$ eV.

III. DISCUSSION

A lightweight MLP operating on frozen MACE embeddings (MACE-dGNN) achieves an element-wise generalization MAE of 0.38 eV, nearly halving the baseline dGNN error of 0.72 eV. Furthermore, MACE-dGNN's element-wise generalization modestly outperforms the zero-shot MACE calculations, MACE-rlx-μ (MAE of 0.51 eV). Following previous works that used machine-learning force fields for defect relaxations [52–55], MACE-rlx-μ replicates the DFT vacancy-energy workflow at reduced cost: defective supercells are constructed and relaxed, total-energy differences are computed, and a reference-state correction is applied. MACE-dGNN operates only on the pristine host structure: it reads the frozen embedding vector at the vacancy site and maps it to a formation energy through a shallow network, requiring neither supercell construction nor relaxation. That MACE-dGNN yields improved accuracy establishes that the pre-readout MACE embedding already encodes the local structural and chemical information relevant to vacancy formation, including the consequences of the atomic relaxation that MACE-dGNN never performs.

Beyond modest improvement from CV error metrics, a natural question arises as to when and why one should choose the embedding-based approach over direct uMLIP calculations. For neutral vacancy formation energies computed at the PBE level of theory, direct MACE relaxation is a viable alternative, although it still requires, as shown here, some existing DFT data on which an optimal μ^{shift} can be fit. A stronger case for the embedding approach is its potential for extensibility. Several defect properties cannot be expressed as energy differences between pristine and defective supercells within a single potential energy surface. Charged defect formation energies depend on the Fermi level position, require band-edge alignment and finite-size corrections [37, 56], and involve electronic degrees of freedom that classical interatomic potentials do not describe; recent ML approaches to this problem operate on structural features rather than total energies. Kiyohara et al. [35] have already demonstrated the ability to train dGNN-style models (i.e., learned from scratch) to predict properties related to charged defects, and we anticipate that MACE-dGNN's fine-tuned embedding approach could further improve generalizability for these types of models. Properties requiring beyond-PBE accuracy, such as formation energies computed with hybrid functionals [12], are accessible to the embedding approach: the MACE embeddings encode PBE-level structural chemistry that correlates with, but does not reproduce, the energet-

ics of the Heyd–Scuseria–Ernzerhof hybrid functional, and a downstream model could in principle be trained on a small set of hybrid-functional calculations to predict such energetics directly. Related work has demonstrated this principle in periodic condensed-phase systems: machine-learning interatomic potentials for liquid water at coupled-cluster and auxiliary-field quantum Monte Carlo accuracy can be trained from as few as 200 high-level energies by initializing from a model pretrained on a cheaper density functional, whereas the same 200 energies are insufficient to train a stable model from scratch. [57] In each of these settings, the embedding approach decouples the representation (provided by the pretrained universal potential) from the prediction target (defined by the task-specific training data), enabling a single set of embeddings to serve multiple downstream tasks.

The MACE-dGNN model benefits from the chemical knowledge encoded during pretraining on millions of diverse structures, a knowledge base far larger than the ∼1,900 vacancy energies available for task-specific training. This is the standard mechanism by which transfer learning improves downstream predictions. The pretrained embeddings are freely available as a community resource, and the relevant practical question is whether they improve downstream predictions for vacancy formation energies; the results presented here show that they do. A zero-hidden-layer linear regression on the 256-dimensional MACE embedding (MACE-LR) achieves an element-wise MAE of 0.55 eV, already lower than the dGNN's 0.72 eV. This indicates that the improvement is driven primarily by the quality of the pretrained representation rather than by the capacity of the downstream model.

Our results contribute to a growing body of work on repurposing pretrained uMLIP representations for downstream property prediction. The "franken" framework of Novelli et al. [31] extracts MACE-MP0 atomic descriptors and could adapt them via kernel methods for surface and defect energies; the HackNIP pipeline of Kim et al. [30] similarly extracts embeddings from pretrained potentials for use in shallow ML models; and the Δ-model approach of Christiansen and Hammer [32] applies corrections to foundation model predictions using the model's own internal representations. Frozen transfer learning, in which lower-layer parameters of a pretrained MACE foundation model are held fixed during fine-tuning, has been shown to reach near-DFT accuracy for reactive surface chemistry and multi-phase alloy datasets using only 10 – 20 % of the data required to train an equivalent model from scratch. [58] Our work shares the core premise of these studies but differs in several respects. Most directly, the concurrent framework of Linton et al. [59] combines a fine-tuned CHGNet

potential with a CGCNN that uses predicted Bader charges as electronic-structure descriptors to predict vacancy formation energies in Ni–Cu–Au–Pd face-centered cubic (FCC) alloys, achieving a root mean square error of 0.06 eV when training on binary and ternary data and predicting in quaternary compositions. While that workflow demonstrates the viability of bypassing DFT entirely for vacancy energetics within a fixed structure type, several distinctions are worth noting.

First, our approach uses frozen MACE embeddings with no parameter updates to the foundation model, whereas Linton et al. fine-tune CHGNet and train a separate Bader charge predictor, creating a multi-stage workflow in which errors can compound. Because the 256-dimensional MACE embedding is the result of message passing optimized to reproduce energies and forces at scale on the OMAT dataset [42], it encodes information correlated with the local electronic environment, including features that scale with formal oxidation state and bonding character; the auxiliary Bader-charge stage of Ref. [59] supplies an analogous signal to a representation (CGCNN atom features) that does not otherwise encode it. Second, Linton et al. test generalization by compositional interpolation within the FCC lattice (binary → ternary → quaternary alloys of the same elements), whereas our element-wise CV withholds all structures containing a given cation, a more stringent test of extrapolation to novel chemistries across multiple structure types. Third, our dataset spans ∼250 oxide compounds encompassing diverse crystal structures (perovskites, spinels, rock salts, rutiles, and others), in contrast to the structurally homogeneous FCC supercells of Ref. [59]. These differences are complementary: the Linton et al. framework is suited to the data-rich, structurally uniform regime of metallic alloys, while the frozen-embedding approach demonstrated here is suited to the data-scarce, structurally diverse regime characteristic of oxide materials discovery.

Beyond uMLIP-based approaches, the most direct prior surrogate model for neutral oxygen vacancy formation energies in oxides is the random-forest model of Kumagai et al. [36], which reports MAE ≈0.34 eV using hand-engineered features on ∼1,500 oxygen vacancy calculations. On a comparable-size dataset of ∼1,100 oxygen vacancy formation energies, MACE-dGNN achieves a structure-wise MAE of 0.33 eV and an element-wise MAE of 0.38 eV (Sec. II C) without manual feature engineering, with the latter providing an additional measurement of generalization across cation chemistry on which hand-engineered models, due to their reliance on element-specific descriptors, are typically not benchmarked.

Our training data spans first-row transition metals (Mn, Fe, Co, Ni) that Kumagai et al. excluded as a methodological constraint of their PBEsol($+U$) workflow [36] but that are central to thermochemical water-splitting candidates [12]; the cation-wise stratification reported here therefore covers the chemistries most relevant to that application.

This work has several limitations. First, our training dataset encompasses $\sim$250 unique oxide structures with nonuniform coverage of elemental space; several cations appear in only one or two binary oxides, limiting the extent to which element-wise generalization can be assessed for those species. Second, we have tested only one family of pretrained models (MACE models trained against OMAT data: `mace-mh-0` for site energies and `mace-omat-0-small` for embeddings); sensitivity to the choice of foundation model, including comparisons with CHGNet, M3GNet, and other MACE variants, remains to be systematically characterized. Third, we restrict attention to neutral vacancies; charged defect formation energies involve additional complexities (Fermi level dependence, finite-size corrections, band-edge alignment) that will require methodological extensions. Finally, the constant-shift correction applied to MACE-sp and MACE-rlx predictions uses a single global offset optimized over the full dataset. We have verified that per-fold shifts are consistent with the global value, but composition-dependent systematic errors, particularly for transition-metal oxides, may persist after a constant correction.

IV. CONCLUSION

We have demonstrated that frozen embeddings from a pretrained MACE universal machine-learning interatomic potential (uMLIP) encode sufficient information about local chemical environments to predict oxygen vacancy formation energies with accuracy exceeding both baseline structure-property models and direct MACE relaxation calculations. Using nested cross-validation with element-wise data splits (the most stringent benchmark for materials discovery), a lightweight multilayer perceptron on 256-dimensional MACE embeddings (MACE-dGNN) achieves a mean absolute error of 0.38 eV, compared to 0.51 eV for the full MACE relaxation workflow and 0.72 eV for the baseline defect graph neural network (dGNN) trained from scratch. The modest advantage of the embedding-based surrogate over direct calculations establishes that pretrained uMLIP representations encode information sufficient to predict vacancy formation energies, including the effects of structural

relaxation, without explicit defect calculations. A decomposition of MACE site energies identifies the neighbor response term Δ_j as the term that accounts for energy differences between pristine bulk and defected supercells in the many-body MACE potential.

Because the embedding approach separates the pretrained representation from the downstream prediction target, it should extend to properties that direct uMLIP energy differences cannot access: charged defect formation energies, vacancy migration barriers, and properties requiring levels of theory beyond the Perdew, Burke, and Ernzerhof functional. The practical significance of improved out-of-sample generalization is illustrated by a thermochemical water-splitting screening in which $\sim$45 % of candidate classifications change relative to the original dGNN predictions, showing the sensitivity of materials discovery workflows to model accuracy in the element-wise extrapolation regime.

The methodology demonstrated here is not specific to oxygen vacancies or to the MACE family of potentials. Pretrained embeddings from any uMLIP can, in principle, serve as transferable features for site-resolved property prediction, provided the foundation model's training data spans the relevant chemical and structural diversity. Systematic benchmarking across foundation models, defect types, and target properties, with consistent use of element-wise or composition-wise splits, will be needed to establish the scope and limits of this approach. As universal potentials continue to improve in accuracy and coverage, the embeddings they produce should serve as general-purpose features for materials property prediction, complementing or replacing task-specific representation learning.

V. METHODS

A. Density functional theory re-validation

To further validate MACE-dGNN screening predictions, we separately computed oxygen vacancy formation energies via density functional theory (DFT) for the fifteen compounds shown in Figure 6f, which are drawn from regions of chemical space sparsely sampled in our training dataset (Figure 6e). The generalized gradient approximation of Perdew, Burke, and Ernzerhof (PBE) was used for these calculations [40], and for compounds containing transition metals (here, Fe, Mo, W), a Hubbard U correction of 3 eV was applied to their d orbitals. The pseudopotentials used are summarized in Table III. As we use the "soft"

oxygen pseudopotential (O_s), which overestimates the bond length of the O_2 molecule, we corrected the energy of molecular oxygen using the molecular binding energy as calculated with the "hard" pseudopotential (O_h), which is added to the energy of an oxygen atom calculated with the soft pseudopotential; this approach results in a reference oxygen chemical potential of 5.04 eV. This approach yields accurate energies for O_2 without requiring a more computationally expensive oxygen pseudopotential for supercell calculations [60]. The O_s pseudopotential allows a plane-wave cutoff energy of 360 eV. Otherwise, total energies are converged to within 10^{-6} eV, and forces are converged to 10 meV Å^{-1}.

TABLE III. Details of VASP pseudopotentials used in this work. All are taken from the VASP version 4.6 repository.

Element	POTCAR	Valence Electron Configuration
Ba	Ba_sv	$4s^2 3p^1$
Ca	Ca_sv	$4s^2 3p^6 5s^2$
Fe	Fe	$4s^1 3d^7$
In	In_d	$5s^2 4d^{10} 5p^1$
K	K_sv	$3s^2 3p^6 4s^1$
Mo	Mo_pv	$4p^6 5s^2 4d^4$
O	O_s	$2s^2 2p^4$
Sn	Sn_d	$5s^2 4d^{10} 5p^2$
Rb	Rb_sv	$4s^2 3p^6 5s^1$
W	W_pv	$4s^2 3d^3$
Zn	Zn	$4s^2 3d^{10}$

The compounds considered in this manner are summarized in Table IV. Also listed are Materials Project identification (MP-ID) numbers, **k**-point meshes used for converging the unit cell, and supercell size, expressed as expansions along the a, b, and c lattice vectors. A single **k**-point is used for all supercell calculations: where bulk compounds are relaxed with a Γ-centered mesh, the Γ point is used; where bulk compounds are relaxed with a Monkhorst–Pack mesh, a single special **k**-point is used. For Fe-containing compounds, antiferromagnetic spin configurations minimize the total energy in all cases tested, and we maintain these spin orderings for all subsequent calculations. Spin polarization is included in all calculations,

but for compounds that do not contain Fe, no initial magnetic ordering is specified. Neutral oxygen vacancy formation energies are then calculated according to Equation (1).

TABLE IV. Compounds considered for additional oxygen vacancy formation energy calculations in this work. The MP-ID numbers are provided, along with the **k**-point meshes used to converge the unit cells, and the supercell sizes used for defect calculations, expressed as expansions of the a, b, and c lattice vectors. For each **k**-point mesh, the table indicates whether it is Γ-centered or Monkhorst–Pack.

Compound	MP-ID	**k**-point mesh	Supercell size
$Ba_3In_2MoO_9$	mp-1228643	Γ, $10 \times 10 \times 6$	$2 \times 2 \times 2$
$Ba_3In_2WO_9$	mp-1228476	Monkhorst–Pack, $8 \times 4 \times 1$	$2 \times 2 \times 1$
$BaMoO_4$	mp-19276	Monkhorst–Pack, $10 \times 10 \times 2$	$2 \times 2 \times 1$
$FeSnO_3$	mp-1178212	Γ, $8 \times 8 \times 8$	$2 \times 2 \times 2$
$FeWO_4$	mp-19421	Γ, $12 \times 10 \times 12$	$3 \times 2 \times 2$
$K_2Sn_3O_7$	mp-1024073	Monkhorst–Pack, $12 \times 2 \times 1$	$3 \times 1 \times 1$
$KBaFeO_3$	mp-18038	Monkhorst–Pack, $8 \times 2 \times 2$	$1 \times 4 \times 2$
$KFeSn_3O_8$	mp-1223574	Γ, $2 \times 12 \times 4$	$2 \times 2 \times 2$
$Rb_2In_4O_7$	mp-27563	Γ, $8 \times 8 \times 2$	$2 \times 2 \times 2$
$Rb_2Sn_3O_7$	mp-2646944	Γ, $1 \times 12 \times 2$	$1 \times 3 \times 1$
$Rb_2W_2O_7$	mp-19144	Γ, $12 \times 2 \times 9$	$3 \times 1 \times 2$
$Rb_2Zn_3O_4$	mp-29606	Γ, $1 \times 6 \times 6$	$1 \times 2 \times 2$
$RbFeO_2$	mp-755536	Monkhorst–Pack, $10 \times 10 \times 6$	$2 \times 2 \times 2$
$RbIn_3O_5$	mp-1209459	Monkhorst–Pack, $8 \times 2 \times 1$	$4 \times 1 \times 1$
Zn_2SnO_4	mp-35493	Monkhorst–Pack, $8 \times 8 \times 6$	$2 \times 2 \times 2$

B. MACE embedding extraction

Embeddings were extracted from the `mace-omat-0-small` model (`https://github.com/ACEsuit/mace-foundations`), a member of the MACE family of equivariant message-passing neural networks [24, 25] pretrained on the Open Materials 2024 dataset [42] with the PBE exchange–correlation head [26]. We selected this model for two reasons: its PBE

training head aligns with the level of DFT theory used in our vacancy formation energy database, and its compact architecture ("small" variant) reduces the risk of overfitting when used to produce fixed input features for downstream models trained on our small dataset of ∼1,900 vacancy energies.

For each host crystal structure in the database, we computed a forward pass of the pretrained MACE model using the Atomic Simulation Environment interface [61]. The 256-dimensional pre-readout feature vector $\boldsymbol{x}_i^{\mathrm{MACE}} \in \mathbb{R}^{256}$ was extracted for each crystallographic site i at the output of the final MACE interaction block (T message-passing steps), before the linear energy readout that produces the scalar site energy E_i^{MACE}. These embeddings were computed once and stored; they serve as fixed (frozen) input features for all downstream models. No MACE parameters are updated during training of MACE-LR or MACE-dGNN.

C. Model architectures

We compare six models spanning a hierarchy of complexity and information access. All models predict the neutral vacancy formation enthalpy $\Delta H_{V_{\mathrm{X},i}}$ for a given crystallographic site i in a host structure.

a. MACE-E_i (scalar site energy baseline). A two-parameter linear regression on the scalar MACE site energy: $\Delta H_{V_{\mathrm{X},i}} = a \cdot E_i^{\mathrm{MACE}} + b$. This model tests whether the scalar energy readout alone correlates with vacancy formation energies.

b. MACE-LR (linear regression on embeddings). A linear regression on the full 256-dimensional embedding vector: $\Delta H_{V_{\mathrm{X},i}} = (\boldsymbol{x}_i^{\mathrm{MACE}})^\top \boldsymbol{\beta}$ (Eq. (4)).

c. Defect graph neural network (baseline, trained from scratch). The defect graph neural network (dGNN), introduced in Ref. [23], adapts the crystal graph convolutional neural network architecture [22] for site-resolved property prediction. The host crystal structure is represented as a graph with atomic nodes and distance-dependent edge features. After $T = 2$ graph convolution layers, the node feature vector corresponding to the vacancy site is extracted and passed through a fully connected readout network. The architecture comprises ∼2,000 learnable parameters, with 8-dimensional node features and 16-dimensional vacancy features. The model is trained from scratch on each cross-validation (CV) fold; full architecture and hyperparameter details are given in Ref. [23] and the supplementary information of Ref. [12].

d. MACE-dGNN (multilayer perceptron on frozen embeddings). A multilayer perceptron (MLP) operating on the frozen 256-dimensional MACE embedding of the vacancy site (Fig. 1c): $\Delta H_{V_{\mathrm{X},i}} = f_{\mathrm{MLP}}(\boldsymbol{x}_i^{\mathrm{MACE}}; \boldsymbol{\theta})$ (Eq. (5)). The MLP has two hidden layers of dimension 64 with rectified linear unit activation functions. The network is trained for a maximum of 1,000 epochs using the Adam optimizer, a learning rate of 0.001, and batch size of 128. Training follows the same nested CV protocol and early-stopping procedure as dGNN (Sec. V E).

e. MACE-sp and MACE-rlx (direct MACE calculations). These are zero-shot models that use MACE as an interatomic potential calculator to compute vacancy formation energies directly, as defined in Equations (6) and (10). For MACE-rlx, both the pristine and defective supercells are relaxed (internal coordinates only, fixed cell parameters) using the limited-memory Broyden–Fletcher–Goldfarb–Shanno optimizer with a force convergence criterion of $0.01\,\mathrm{eV\,\AA^{-1}}$. For MACE-sp, total energies are computed at the DFT-relaxed geometries without further optimization. The O_2 reference energy is taken from a MACE relaxation of the isolated molecule. A constant shift $\mu_{\mathrm{O}}^{\mathrm{shift}}$ is applied to MACE predictions to correct for the difference in reference state from the DFT data (see Sec. V D).

MACE-rlx and MACE-dGNN differ in their input requirements: MACE-rlx requires interatomic potential relaxation of both pristine bulk and defective supercells, while MACE-dGNN takes only the pristine host structure as input.

D. Direct MACE vacancy formation energy workflow

The MACE-sp and MACE-rlx workflows compute vacancy formation energies by evaluating total energy differences between pristine and defective supercells using the pretrained MACE potential, as defined in Equations (6) and (10). Because MACE and DFT employ different energy reference conventions and exchange–correlation treatments, the raw MACE formation energies are systematically offset from the DFT values. To remove this systematic offset, we apply a constant shift $\mu_{\mathrm{O}}^{\mathrm{shift}}$ defined as the value minimizing the mean absolute error (MAE) of MACE predictions with respect to the DFT reference data over the full dataset:

$$\mu_{\mathrm{O}}^{\mathrm{shift}} = \arg\min_{\mu} \frac{1}{N} \sum_{i=1}^{N} \left| \Delta H_{V_{\mathrm{O},i}}^{\mathrm{MACE}} + \mu - \Delta H_{V_{\mathrm{O},i}}^{\mathrm{DFT}} \right|. \tag{13}$$

We find this shift to be $0.91\,\mathrm{eV}$ for MACE-rlx and $-0.12\,\mathrm{eV}$ for MACE-sp. To ensure

that this correction does not introduce information leakage in the CV evaluation, we verified that shifts computed independently within each outer training fold are consistent with the global value. We also report unshifted MAE values in the supplementary information to enable assessment of the raw MACE prediction quality.

E. Cross-validation strategy

MACE-LR, dGNN, and MACE-dGNN were all trained and tested with nested (K, L)-fold CV, using identical splits across models to ensure consistency. For structure-wise splits with $(|K|, |L|) = (10, 10)$, all defects from a given structure are randomly assigned to either the outer train or test split. For element-wise splits with $(|K|, |L|) = (18, 10)$, all defects from any structure containing the targeted test cation (Al, Ba, Ca, Co, Cr, Fe, In, La, Mg, Mn, Nb, Ni, Pr, Sn, Sr, Ti, V, or Y) are held out for the outer test set (with the exception of any binary oxides, which are always assigned to the outer train set). For both outer split strategies, random $|L| = 10$-fold inner train/test splits are generated.

Taking the set of all $|L| = 10$ inner fold models' predictions on a given outer test sample, the final surrogate model's prediction is the ensemble's average,

$$\Delta H_{V_\mathrm{X}}^{\mathrm{model}} = 1/|L| \sum_{l \in L} \Delta H_{V_\mathrm{X},l}, \tag{14}$$

and the metric for uncertainty is the sample standard deviation across the L-fold ensemble,

$$\sigma(H_{V_\mathrm{X}}) = \sqrt{1/(L-1) \sum_{l \in L} \left| \Delta H_{V_\mathrm{X}}^{\mathrm{model}} - \Delta H_{V_\mathrm{X},l} \right|^2}. \tag{15}$$

Expected model performance for both CV splitting scenarios is assessed by MAE and coefficient of determination (R^2). Two scenarios are considered given our dataset's imbalance (specifically in the element-wise splits), reflected in both the uneven distribution of $\Delta H_{V_\mathrm{X}}^{\mathrm{DFT}}$ target values and the uneven coverage of chemical space. We quantify a global MAE averaged across all N outer test-set predictions in the dataset,

$$\mathrm{MAE} = 1/|N| \sum_{i \in N} \left| \Delta H_{V_\mathrm{X},i}^{\mathrm{model}} - \Delta H_{V_\mathrm{X},i}^{\mathrm{DFT}} \right|. \tag{16}$$

Alternatively, we quantify the unweighted mean of the per-fold MAEs,

$$\langle \{\mathrm{MAE}\}_K \rangle = 1/|K| \sum_{k \in K} \mathrm{MAE}_k. \tag{17}$$

The same stratification of test predictions is used to determine R^2 and $\langle\{R^2\}_K\rangle$. For the zero-shot models MACE-E_i and MACE-rlx, MAE, $\langle\{\text{MAE}\}_K\rangle$, R^2, and $\langle\{R^2\}_K\rangle$ are computed from their predictions on the same outer splits.

F. Embedding analysis

Dimensionality reduction of the 256-dimensional MACE embeddings was performed using Uniform Manifold Approximation and Projection [62, 63] as implemented in the umap-learn Python package (0.5.12). Two-component projections (Fig. 2b–c) were computed with `n_neighs_umap` = 10, `n_components` = 2, and `min_dist` = 0.99.

ACKNOWLEDGMENTS

We thank Rose Cersonsky, Ethan Deutsch, and Christian Jorgensen (University of Wisconsin–Madison) for discussions on interpreting MLIP embeddings in the context of point-defect formation energies. This work was supported by the U.S. Department of Energy's Office of Critical Minerals and Energy Innovation (CMEI) under the Alternative Fuels and Feedstocks Office's award number DE-EE0010733 and by the Laboratory Directed Research and Development (LDRD) program at Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration (DOE/NNSA) under contract DE-NA0003525. This written work is authored by an employee of NTESS. The employee, not NTESS, owns the right, title and interest in and to the written work and is responsible for its contents. Any subjective views or opinions that might be expressed in the written work do not necessarily represent the views of the U.S. Government. The publisher acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this written work or allow others to do so, for U.S. Government purposes. The DOE will provide public access to results of federally sponsored research in accordance with the DOE Public Access Plan. Part of this work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract

DE-AC52-07NA27344. The National Laboratory of the Rockies (NLR) is operated for the DOE under Contract No. DE-AC36-08GO28308. A portion of the research was performed using computational resources sponsored by the Department of Energy and located at the National Laboratory of the Rockies.

DATA AVAILABILITY

The DFT vacancy formation enthalpy data on which this study is based are available at Zenodo (Ref. [64], `https://zenodo.org/records/8087871`) and in the supporting files of Ref. [41]. The training data (CIF files, defect sites, and vacancy formation energies), nested train/test splits, trained model weights, and MACE-dGNN test predictions will be provided in the supporting files upon publication. Materials Project structures used as inputs for thermochemical water-splitting screening are accessible via the Materials Project API (`https://materialsproject.org`).

CODE AVAILABILITY

The d^2GNN training code, which includes the dGNN model, is available at `https://github.com/sandialabs/d2gnn`. A Python script that runs the MACE-dGNN training will be provided in the supporting files upon publication. Pretrained MACE foundation models (`mace-mh-0`, `mace-omat-0-small`) are publicly available from the MACE Foundations repository (`https://github.com/ACEsuit/mace-foundations`).

AUTHOR CONTRIBUTIONS

Conceptualization: M.D.W., R.B.W. **Methodology**: M.D.W., R.B.W., S.P. **Software**: M.D.W., S.P. **Validation**: M.D.W., S.P. **Formal analysis**: M.D.W., R.B.W. **Investigation**: S.P., A.J.E.R., S.L. **Data curation**: M.D.W. **Writing – original draft**: R.B.W., M.D.W. **Writing – review and editing**: all authors. **Visualization**: M.D.W., S.P. **Supervision**: M.D.W., R.B.W., J.B.V., S.L. **Project administration**: M.D.W., R.B.W. **Funding acquisition**: M.D.W., R.B.W., J.B.V., S.L.

COMPETING INTERESTS

The authors declare no competing interests.

---

[1] A. Muñoz-García, A. M. Ritzmann, M. Pavone, J. Keith, and E. Carter, Oxygen transport in perovskite-type solid oxide fuel cell materials: insights from quantum mechanics., Acc. Chem. Res. **47**, 3340 (2014).

[2] H. Cai, C. Xia, X. Wang, W. Dong, H. Xiao, D. Zheng, H. Wang, and B. Wang, Diverse functions of oxygen vacancies for oxygen ion conduction, ACS Appl. Energy Mater. **5**, 11122 (2022).

[3] A. Rowberg, H. S. Slomski, N. Kim, N. A. Strange, B. P. Gorman, S. Shulda, D. Ginley, K. E. Kweon, and B. C. Wood, Impact of sr-containing secondary phases on oxide conductivity in solid-oxide electrolyzer cells, Chem. Mater. **36**, 6464 (2024).

[4] D. E. Matkin, I. A. Starostina, M. B. Hanif, and D. A. Medvedev, Revisiting the ionic conductivity of solid oxide electrolytes: A technical review, J. Mater. Chem. A **12**, 25696 (2024).

[5] K. Yu, L.-L. Lou, S. Liu, and W. Zhou, Asymmetric oxygen vacancies: the intrinsic redox active sites in metal oxide catalysts, Adv. Sci. **7**, 1901970 (2019).

[6] H. Idriss, Oxygen vacancies role in thermally driven and photon driven catalytic reactions, Chem Catal. **2**, 1549 (2022).

[7] Y. Han, J. Xu, W. Xie, Z. Wang, and P. Hu, Comprehensive study of oxygen vacancies on the catalytic performance of zno for co/h2 activation using machine learning-accelerated first-principles simulations, ACS Catal. **13**, 5104 (2023).

[8] X. Wang, S. Xue, M. Huang, W. Lin, Y. Hou, Z. Yu, M. Anpo, J. C. Yu, J. Zhang, and X. Wang, Pressure-induced engineering of surface oxygen vacancies on metal oxides for heterogeneous photocatalysis., J. Am. Chem. Soc. **147**, 4945 (2025).

[9] J. Yang, M. Pickett, X. Li, D. Ohlberg, D. Stewart, and R. S. Williams, Memristive switching mechanism for metal/oxide/metal nanodevices., Nat. Nanotechnol. **3**, 429 (2008).

[10] R. Schmitt, J. Spring, R. Korobko, and J. Rupp, Design of oxygen vacancy configuration for memristive systems., ACS Nano **11**, 8881 (2017).

[11] R. L. Mártir, E. Jagla, D. Rubi, and M. J. Sánchez, Oxygen vacancies driven filamentary resistive switching in oxide-based memristive devices, J. Phys. D: Appl. Phys. **58**, 325306 (2025).

[12] T. C. Douglas, M. J. Dzara, A. J. E. Rowberg, K. A. King, M. Syrigou, N. A. Strange, R. T. Bell, A. Goyal, P.-W. Guan, R. B. Wexler, J. B. Varley, T. Ogitsu, S. Lany, A. H. McDaniel, S. R. Bishop, and M. D. Witman, Large-scale experimental validation of thermochemical water-splitting oxides discovered by defect graph neural networks, Mater. Horiz. **13**, 829 (2026).

[13] J. Buckeridge, C. R. A. Catlow, M. R. Farrow, A. J. Logsdail, D. O. Scanlon, T. W. Keal, P. Sherwood, S. M. Woodley, A. A. Sokol, and A. Walsh, Deep vs shallow nature of oxygen vacancies and consequent n-type carrier concentrations in transparent conducting oxides, Phys. Rev. Mater. **2**, 054604 (2018).

[14] J. Chen, N. Bogdanov, D. Usvyat, W. Fang, A. Michaelides, and A. Alavi, The color center singlet state of oxygen vacancies in TiO_2, J. Chem. Phys. **153**, 204704 (2020).

[15] Y. Chen, M. E. Turiansky, and C. G. Van de Walle, First-principles study of quantum defect candidates in beryllium oxide, Phys. Rev. B **106**, 174113 (2022).

[16] A. Bayon, A. de la Calle, E. B. Stechel, and C. Muhich, Operational limits of redox metal oxides performing thermochemical water splitting, Energy Technol. **10**, 2100222 (2021).

[17] R. B. Wexler, E. B. Stechel, and E. A. Carter, Materials design directions for solar thermochemical water splitting, in *Solar Fuels*, edited by N. D. Sankir and M. Sankir (Scrivener Publishing, Beverly, MA, 2023) Chap. 1, pp. 3–64.

[18] S. Lany, Chemical potential analysis as an alternative to the van't Hoff method: Hypothetical limits of solar thermochemical hydrogen, Journal of the American Chemical Society **146**, 14114 (2024).

[19] 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, Commentary: The materials project: A materials genome approach to accelerating materials innovation, APL Materials **1**, 011002 (2013).

[20] M. K. Horton, P. Huck, R. X. Yang, J. M. Munro, S. Dwaraknath, A. M. Ganose, R. S. Kingsbury, M. Wen, J. X. Shen, T. S. Mathis, A. D. Kaplan, K. Berket, J. Riebesell, J. George, A. S. Rosen, E. W. C. Spotte-Smith, M. J. McDermott, O. A. Cohen, A. Dunn, M. C. Kuner, G.-M. Rignanese, G. Petretto, D. Waroquiers, S. M. Griffin, J. B. Neaton, D. C. Chrzan,

M. Asta, G. Hautier, S. Cholia, G. Ceder, S. P. Ong, A. Jain, and K. A. Persson, Accelerated data-driven materials science with the Materials Project, Nat. Mater. **24**, 1522 (2025).

[21] M. D. Witman and P. Schindler, Matfold: systematic insights into materials discovery models' performance through standardized cross-validation protocols, Digital Discovery **4**, 625 (2025).

[22] 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).

[23] M. D. Witman, A. Goyal, T. Ogitsu, A. H. McDaniel, and S. Lany, Defect graph neural networks for materials discovery in high-temperature clean-energy applications, Nat. Comput. Sci. **3**, 675 (2023).

[24] I. Batatia, D. P. Kovacs, G. N. C. Simm, C. Ortner, and G. Csanyi, MACE: Higher order equivariant message passing neural networks for fast and accurate force fields, in *Advances in Neural Information Processing Systems*, edited by A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho (2022).

[25] I. Batatia, S. Batzner, D. P. Kovács, A. Musaelian, G. N. C. Simm, R. Drautz, C. Ortner, B. Kozinsky, and G. Csányi, The design space of e(3)-equivariant atom-centred interatomic potentials, Nat. Mach. Intell. **7**, 56 (2025).

[26] I. Batatia, P. Benner, Y. Chiang, A. M. Elena, D. P. Kovács, J. Riebesell, X. R. Advincula, M. Asta, M. Avaylon, W. J. Baldwin, F. Berger, N. Bernstein, A. Bhowmik, F. Bigi, S. M. Blau, V. Cărare, M. Ceriotti, S. Chong, J. P. Darby, S. De, F. Della Pia, V. L. Deringer, R. Elijošius, Z. El-Machachi, E. Fako, F. Falcioni, A. C. Ferrari, J. L. A. Gardner, M. J. Gawkowski, A. Genreith-Schriever, J. George, R. E. A. Goodall, J. Grandel, C. P. Grey, P. Grigorev, S. Han, W. Handley, H. H. Heenen, K. Hermansson, C. H. Ho, S. Hofmann, C. Holm, J. Jaafar, K. S. Jakob, H. Jung, V. Kapil, A. D. Kaplan, N. Karimitari, J. R. Kermode, P. Kourtis, N. Kroupa, J. Kullgren, M. C. Kuner, D. Kuryla, G. Liepuoniute, C. Lin, J. T. Margraf, I.-B. Magdău, A. Michaelides, J. H. Moore, A. A. Naik, S. P. Niblett, S. W. Norwood, N. O'Neill, C. Ortner, K. A. Persson, K. Reuter, A. S. Rosen, L. A. M. Rosset, L. L. Schaaf, C. Schran, B. X. Shi, E. Sivonxay, T. K. Stenczel, C. Sutton, V. Svahn, T. D. Swinburne, J. Tilly, C. van der Oord, S. Vargas, E. Varga-Umbrich, T. Vegge, M. Vondrák, Y. Wang, W. C. Witt, T. Wolf, F. Zills, and G. Csányi, A foundation model for atomistic materials chemistry, J. Chem. Phys. **163**, 184110 (2025).

[27] B. Deng, P. Zhong, K. Jun, J. Riebesell, K. Han, C. J. Bartel, and G. Ceder, Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling, Nat. Mach. Intell. **5**, 1031 (2023).

[28] C. Chen and S. P. Ong, A universal graph deep learning interatomic potential for the periodic table, Nat. Comput. Sci. **2**, 718 (2022).

[29] D. Zhang, H. Bi, F.-Z. Dai, W. Jiang, X. Liu, L. Zhang, and H. Wang, Pretraining of attention-based deep learning potential model for molecular simulation, npj Comput. Mater. **10**, 94 (2024).

[30] S. Y. Kim, Y. J. Park, and J. Li, Leveraging neural network interatomic potentials for a foundation model of chemistry (2025), arXiv:2506.18497 [cond-mat.mtrl-sci].

[31] P. Novelli, G. Meanti, P. J. Buigues, L. Rosasco, M. Parrinello, M. Pontil, and L. Bonati, Fast and fourier features for transfer learning of interatomic potentials, npj Comput. Mater. **11**, 293 (2025).

[32] M. V. Christiansen and B. Hammer, Δ-model correction of foundation model based on the model's own understanding, J. Chem. Phys. **162**, 184701 (2025).

[33] B. Deng, Y. Choi, P. Zhong, J. Riebesell, S. Anand, Z. Li, K. Jun, K. A. Persson, and G. Ceder, Systematic softening in universal machine learning interatomic potentials, npj Comput. Mater. **11**, 9 (2025).

[34] B. Focassio, L. P. M. Freitas, and G. R. Schleder, Performance assessment of universal machine learning interatomic potentials: Challenges and directions for materials' surfaces., ACS Appl. Mater. Interfaces **17**, 13111 (2025).

[35] S. Kiyohara, C. Shibui, S. Bae, and Y. Kumagai, Machine-learning prediction of charged-defect formation energies from crystal structures, Physical Review Letters **135**, 246101 (2025).

[36] Y. Kumagai, N. Tsunoda, A. Takahashi, and F. Oba, Insights into oxygen vacancies from high-throughput first-principles calculations, Physical Review Materials **5**, 123803 (2021).

[37] C. Freysoldt, B. Grabowski, T. Hickel, J. Neugebauer, G. Kresse, A. Janotti, and C. G. Van de Walle, First-principles calculations for point defects in solids, Rev. Mod. Phys. **86**, 253 (2014).

[38] L. Way, C. D. Spataru, R. E. Jones, D. R. Trinkle, A. J. E. Rowberg, J. B. Varley, R. B. Wexler, C. M. Smyth, T. C. Douglas, S. R. Bishop, E. J. Fuller, A. H. McDaniel, S. Lany, and M. D. Witman, Defect diffusion graph neural networks for materials discovery in high-temperature energy applications, Chemistry of Materials **37**, 6473 (2025).

[39] R. Devi, K. T. Butler, and G. S. Gautam, Leveraging transfer learning for accurate estimation of ionic migration barriers in solids, npj Computational Materials 10.1038/s41524-026-01972-8 (2026).

[40] J. P. Perdew, K. Burke, and M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. **77**, 3865 (1996).

[41] R. B. Wexler, G. S. Gautam, E. B. Stechel, and E. A. Carter, Factors governing oxygen vacancy formation in oxide perovskites, J. Am. Chem. Soc. **143**, 13212 (2021).

[42] L. Barroso-Luque, M. Shuaibi, X. Fu, B. M. Wood, M. Dzamba, M. Gao, A. Rizvi, C. L. Zitnick, and Z. W. Ulissi, Open materials 2024 (OMat24) inorganic materials dataset and models, arXiv preprint arXiv:2410.12771 10.48550/arXiv.2410.12771 (2024).

[43] J. S. Smith, B. Nebgen, N. Lubbers, O. Isayev, and A. E. Roitberg, Less is more: Sampling chemical space with active learning, J. Chem. Phys. **148**, 241733 (2018).

[44] J. Vandermause, Y. Xie, J. S. Lim, C. J. Owen, and B. Kozinsky, Active learning of reactive bayesian force fields applied to heterogeneous catalysis dynamics of H/Pt, Nat. Commun. **13**, 5183 (2022).

[45] J. Qi, T. W. Ko, B. C. Wood, T. A. Pham, and S. P. Ong, Robust training of machine learning interatomic potentials with dimensionality reduction and stratified sampling, npj Computational Materials **10**, 43 (2024).

[46] M. Kulichenko, B. Nebgen, N. Lubbers, J. S. Smith, K. Barros, A. E. A. Allen, A. Habib, E. Shinkle, N. Fedik, Y. W. Li, R. A. Messerly, and S. Tretiak, Data generation for machine learning interatomic potentials and beyond, Chemical Reviews **124**, 13681 (2024).

[47] A. M. Deml, A. M. Holder, R. P. O'Hayre, C. B. Musgrave, and V. Stevanović, Intrinsic material properties dictating oxygen vacancy formation energetics in metal oxides, J. Phys. Chem. Lett. **6**, 1948 (2015).

[48] V. Stevanović, S. Lany, X. Zhang, and A. Zunger, Correcting density functional theory for accurate predictions of compound enthalpies of formation: Fitted elemental-phase reference energies, Phys. Rev. B **85**, 115104 (2012).

[49] A. Sharan and S. Lany, Computational discovery of stable and metastable ternary oxynitrides, J. Chem. Phys. **154**, 234706 (2021).

[50] A. A. Emery, J. E. Saal, S. Kirklin, V. I. Hegde, and C. Wolverton, High-throughput computational screening of perovskites for thermochemical water splitting applications, Chem. Mater.

28, 5621 (2016).

[51] S. Lany, Communication: The electronic entropy of charged defect formation and its impact on thermochemical redox cycles, J. Chem. Phys. **148**, 071101 (2018).

[52] M. H. Rahman, M. Biswas, and A. Mannodi-Kanakkithodi, Understanding defect-mediated ion migration in semiconductors using atomistic simulations and machine learning, ACS Materials Au **4**, 557 (2024).

[53] A. Mannodi-Kanakkithodi, X. Xiang, L. Jacoby, R. Biegaj, S. T. Dunham, D. R. Gamelin, and M. K. Chan, Universal machine learning framework for defect predictions in zinc blende semiconductors, Patterns **3**, 100450 (2022).

[54] I. Mosquera-Lois, S. R. Kavanagh, A. M. Ganose, and A. Walsh, Machine-learning structural reconstructions for accelerated point defect calculations, npj Computational Materials **10**, 121 (2024).

[55] S. R. Kavanagh, Identifying split vacancy defects with machine-learned foundation models and electrostatics, Journal of Physics: Energy **7**, 045002 (2025).

[56] S. Lany and A. Zunger, Accurate prediction of defect properties in density functional supercell calculations, Modelling and simulation in materials science and engineering **17**, 084002 (2009).

[57] M. S. Chen, J. Lee, H.-Z. Ye, T. C. Berkelbach, D. R. Reichman, and T. E. Markland, Data-efficient machine learning potentials from transfer learning of periodic correlated electronic structure methods: Liquid water at AFQMC, CCSD, and CCSD(T) accuracy, Journal of Chemical Theory and Computation **19**, 4510 (2023).

[58] M. Radova, W. G. Stark, C. S. Allen, R. J. Maurer, and A. P. Bartók, Fine-tuning foundation models of materials interatomic potentials with frozen transfer learning, npj Computational Materials **11**, 237 (2025).

[59] N. Linton, P. Singh, and D. S. Aidhy, Framework to completely bypass expensive DFT calculations via graph neural networks for vacancy formation energy predictions in FCC high entropy alloys, npj Comput. Mater. 10.1038/s41524-026-02037-6 (2026), article in Press.

[60] H. Peng, D. O. Scanlon, V. Stevanovic, J. Vidal, G. W. Watson, and S. Lany, Convergence of density and hybrid functional defect calculations for compound semiconductors, Phys. Rev. B **88**, 115201 (2013).

[61] 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).

[62] L. McInnes, J. Healy, and J. Melville, UMAP: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 10.48550/arXiv.1802.03426 (2018).

[63] J. Healy and L. McInnes, Uniform manifold approximation and projection, Nat. Rev. Methods Primers **4**, 82 (2024).

[64] M. Witman, A. Goyal, T. Ogitsu, A. H. McDaniel, and S. Lany, A database of vacancy formation enthalpies for materials discovery (0.0.1) [data set], Zenodo , 10.5281/zenodo.8087871 (2023).