arXiv:2611.15297v1 [cond-mat.mtrl-sci] 11 Nov 2026

Periodic Trends in Oxygen Vacancy Formation Energies Across Binary Main-Group Metal Oxides

Manish Kumar[1], Bridget S. Ferris[1], Avital Isakov[1,2], and Robert B. Wexler*[1]

[1]Department of Chemistry and Institute of Materials Science and Engineering, Washington University in St. Louis, St. Louis, MO 63130, USA
[2]Present address: Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, CA 90095, USA

*Email: wexler@wustl.edu

Abstract

The function of oxide materials, from solid-oxide fuel cells to solar thermochemical fuel production, depends on their oxygen vacancies. Their formation energetics have been mapped systematically for ternary perovskites, but reported values for the binary oxides, their conceptual and thermochemical building blocks, remain scattered across studies with differing computational methods. Here we report a systematic r^{2}SCAN density functional theory survey of neutral oxygen vacancy formation energies across 39 binary main-group metal oxide polymorphs spanning groups 1, 2, 13, 14, and 15, using the ShakeNBreak method to locate reconstructed defect ground states. The formation energies span 8 eV and decrease down each group and, for most multivalent cations, with increasing oxidation state. Both trends follow the reducibility of the cation. An interpretable crystal feature model is fit to these energies; its inputs are properties of the pristine bulk, with the reduction potential alternatively taken from tabulated thermochemistry. Retaining only its two dominant descriptors, the crystal reduction potential and the band gap, the model reproduces the formation energies with a held-out mean absolute error of 0.31 eV, lower than that of its four-descriptor parent model for transition-metal perovskites. The bond-breaking and stability terms of the full Born–Haber fit prove statistically null for the binary oxides, and the largest residuals identify the hosts where electron accommodation departs from the cation-reduction picture that the descriptors encode. The dataset provides a benchmark for machine-learning interatomic potentials, and the model estimates vacancy formation energies of main-group oxides from bulk properties alone.

Introduction

Oxygen vacancies in metal oxides mediate performance across a broad range of energy technologies. In solid oxide fuel cells, vacancies provide pathways for bulk oxide ion transport and determine the rate of surface oxygen exchange,[1,2] and in solar thermochemical cycles, the vacancy formation energy determines the operating temperatures and thermodynamic efficiency of hydrogen production.[2–4] Oxygen vacancies are also implicated in the n-type conductivity of transparent conducting oxides, where their deep versus shallow character has been debated,[5] in the formation of catalytically active sites for oxidation reactions,[6] and in the degradation of lithium-ion battery cathodes through oxygen loss.[4] A systematic mapping of vacancy formation energetics across oxide chemical space nevertheless remains incomplete.

The energy required to form a neutral oxygen vacancy can be understood through an extended Born–Haber cycle involving two principal contributions.[1,2] First, the metal–oxygen bonds coordinating the vacancy site must be broken, an energetic cost that scales with the strength of local M–O bonding. Second, the two electrons formerly associated with the lattice O^{2-} ion must be accommodated elsewhere in the structure, typically by reducing neighboring cations or, in metallic systems, by populating states near the Fermi level.

The interplay between these bond-breaking and electron-accommodation terms explains why vacancy formation energies correlate with both thermodynamic quantities such as the oxide formation enthalpy and electronic structure features including the band gap and the position of the oxygen 2p band.[1,7] Building on this physical picture, Wexler et al.[2] decomposed the formation energy into contributions from crystal bond dissociation energies, crystal reduction potentials, band gaps, and thermodynamic stability, achieving mean absolute errors of approximately 0.45 eV for perovskite oxides stable at ambient conditions. Linderälv et al.[5] demonstrated that the charge transition levels of oxygen vacancies, when aligned to the vacuum level, exhibit small variation across wide-band-gap binary and ternary oxides, pointing to underlying universality in the electronic structure of the defect state itself. These conceptual advances notwithstanding, translating physical insight into broadly predictive models has proven challenging, in part because the relative importance of local descriptors (properties of the immediate coordination environment of the vacant oxygen) and global descriptors (properties of the host crystal as a whole) varies with crystal structure and composition.

High-throughput computational studies have mapped vacancy formation energies across increasingly large regions of oxide chemical space. Emery and Wolverton[8] computed formation energies for 5329 ABO_3 perovskites. Kumagai et al.[9] performed automated point-defect calculations on 937 oxides (approximately 6000 vacancy calculations, three charge states at each symmetry-inequivalent oxygen site) and trained random forest models with mean absolute errors of 0.27–0.44 eV depending on charge state, and Baldassarri et al.[4] extended this approach to over 2500 vacancy calculations across more than 1000 compounds with comparable accuracy. Graph neural networks have also been applied to vacancy prediction: Witman et al.[10] developed a defect graph neural network that learns directly from crystal graphs without hand-engineered features, and Choudhary and Sumpter[11] demonstrated that a graph neural network energy model trained only on defect-free structures can extrapolate to vacancy formation energies, albeit with reduced accuracy.

Despite this progress, the existing computational literature exhibits a pronounced compositional bias: ternary and quaternary oxides (perovskites in particular) dominate, and the main-group elements of groups 1, 2, 13, 14, and 15 appear sporadically across these datasets but have not been surveyed across all known polymorphs of their binary oxides. Binary oxides allow structure–property relationships to be examined most directly: with only one cation species, variations in vacancy formation energy across polymorphs can be attributed to differences in local coordination, bond lengths, and electronic structure rather than to compositional complexity. Prior descriptor-based models have also used the thermochemistry of binary oxides (their formation enthalpies, cation reduction energetics, and vacancy formation energies) as inputs for predicting vacancy behavior in multinary compounds,[2,4] and the solid-state electrochemical series of Mueller et al.[12] underscores the value of solid-state, rather than aqueous, reference data for organizing redox behavior. Establishing periodic trends in vacancy formation energies across binary oxide chemical space (analogous to the periodic trends in ionization energies or electron affinities that organize atomic behavior) would provide both physical insight and a foundation for rational defect engineering in more complex systems.

Here we report a systematic computational survey of neutral oxygen vacancy formation energies in binary oxides of the main-group metals spanning groups 1 (Li, Na, K, Rb, Cs), 2 (Be, Mg, Ca, Sr, Ba), 13 (Al, Ga, In, Tl), 14 (Sn, Pb), and 15 (Bi). We consider every polymorph of these oxides that the Materials Project[13] lists with an experimental structure from the Inorganic Crystal Structure Database (ICSD),[14] apart from one Bi_2O_3 entry whose structure contains O–O pairs (see Polymorph Selection), and compute the formation energy of every symmetry-inequivalent oxygen vacancy site using density functional theory with the r^{2}SCAN meta-generalized-gradient functional. This choice aligns with the current Materials Project computational framework and extends r^{2}SCAN defect calculations, which have so far targeted individual hosts,[15] to a high-throughput dataset.

We employ the ShakeNBreak approach for defect structure optimization, which samples a diverse set of initial distortions to identify low-energy reconstructed configurations that conventional local relaxation often misses.[16,17] This methodology guards against the systematic overestimation of formation energies that can occur when calculations become trapped in metastable high-symmetry configurations, a source of error that prior high-throughput defect studies, which relaxed each vacancy from a single starting configuration, have not systematically addressed.[4,9,10]

This work makes three principal contributions. First, we adapt the interpretable crystal feature model of Wexler et al.[2] to the binary oxides. Of its four physically motivated descriptors (the crystal bond dissociation energy, the crystal reduction potential, the band gap, and the energy above the convex hull), the reduction potential and the band gap prove sufficient, and the largest residuals of the model identify hosts in which

electron accommodation departs from the cation reduction that the descriptors encode. Second, the dataset maps vacancy formation energies systematically across the binary oxides of these metal groups and establishes quantitative periodic trends in them: across hosts, the formation energies follow cation reducibility and oxidation state; within a host, they follow the local bond geometry of the vacant site. Third, we offer this dataset as a benchmark resource for machine-learning interatomic potentials, for which reference data spanning simple binary systems are needed to assess transferability and identify failure modes.

Methods

Polymorph Selection and Training Data Curation

The dataset comprises 66 material–polymorph combinations of 27 nominal binary oxide compositions of the 17 main-group metals, in all seven crystal systems and 36 space groups (Figure 1). One further Bi_2O_3 entry of the Materials Project (mp-558953) was not computed, because it is not a binary oxide of O^{2-}: a third of its oxygen atoms form O–O pairs 1.52 Å apart. Because the crystal feature model (CFM)[2] is best suited to hosts that are at least metastable under the near-ambient conditions of typical synthesis and operation, the training set was curated from the experimental provenance of each polymorph in the ICSD.[14] Every ICSD collection code that the Materials Project lists was retrieved (467 codes in total), and each polymorph was classified according to whether at least one experimentally determined entry corresponds to near-ambient conditions, defined here as a temperature within 273–313 K and a pressure $\leq$0.2 MPa (classification rules in the Supporting Information).

Of the 66 polymorphs evaluated, 39 were retained for CFM training, and 27 were excluded, in five classes (Figure 1): ten phases that exist only under high hydrostatic pressure; five records that are not binary oxides, such as oxyhydroxides recorded without their hydrogen atoms; nine unstable idealized cells, which are dynamically unstable in the primitive cell or in the vacancy supercell or do not reproduce the recorded structure; two earlier structural models of phases retained under another space group; and one model of a non-bulk layer. The high-pressure phases are excluded because the descriptors, like the formation energies, are evaluated at zero pressure, at which these phases are at best metastable; at the pressures where the phases are stable, the vacancy formation enthalpy acquires a pressure–volume term, the product of the pressure and the vacancy relaxation volume, that none of the descriptors represents. For the records that are not binary oxides and for the unstable idealized cells, no vacancy formation energy of a distinct binary-oxide phase can be defined. The evidence and the structural references for each excluded polymorph are given in the Supporting Information (SI Tables S2 and S4).

The retained training set comprises 39 polymorphs spanning 24 unique compositions. Enumerated with the PYMATGEN vacancy generator on the relaxed unperturbed cell of each, these polymorphs present 72 symmetry-inequivalent oxygen sites, at each of which a vacancy formation energy was computed. The sites span twofold through eightfold oxygen coordination (SI Figure S2), in contrast to the fixed oxygen environments of the perovskite lattice, so the bond-dissociation descriptor defined below varies from site to site through the coordination number. The retained polymorphs that have ICSD entries outside the near-ambient window are discussed in the Supporting Information.

Density Functional Theory Calculations

All density functional theory (DFT) calculations were performed with the Vienna *Ab initio* Simulation Package (VASP, version 6.4)[18] using the projector augmented-wave (PAW) method[19,20] and the r^{2}SCAN meta-generalized-gradient-approximation (meta-GGA) functional,[21,22] adopted for consistency with the r^{2}SCAN thermodynamic data of the Materials Project.[13,23,24] Input files were prepared with PYMATGEN[25] following the Materials Project r^{2}SCAN relaxation set,[23] which applies neither a Hubbard-U correction nor hybrid exchange. The plane-wave basis was truncated at a kinetic-energy cutoff of 520 eV, no symmetry was imposed in any calculation, the oxide calculations were spin-polarized, and geometries were optimized until all interatomic forces fell below 0.03 eV Å^{-1}. The pseudopotentials, the remaining electronic settings, the Brillouin-zone sampling of the bulk cells, and the elemental-metal reference calculations are given in the Supporting Information. SI Figure S1 summarizes the full workflow from source structure to fitted model.

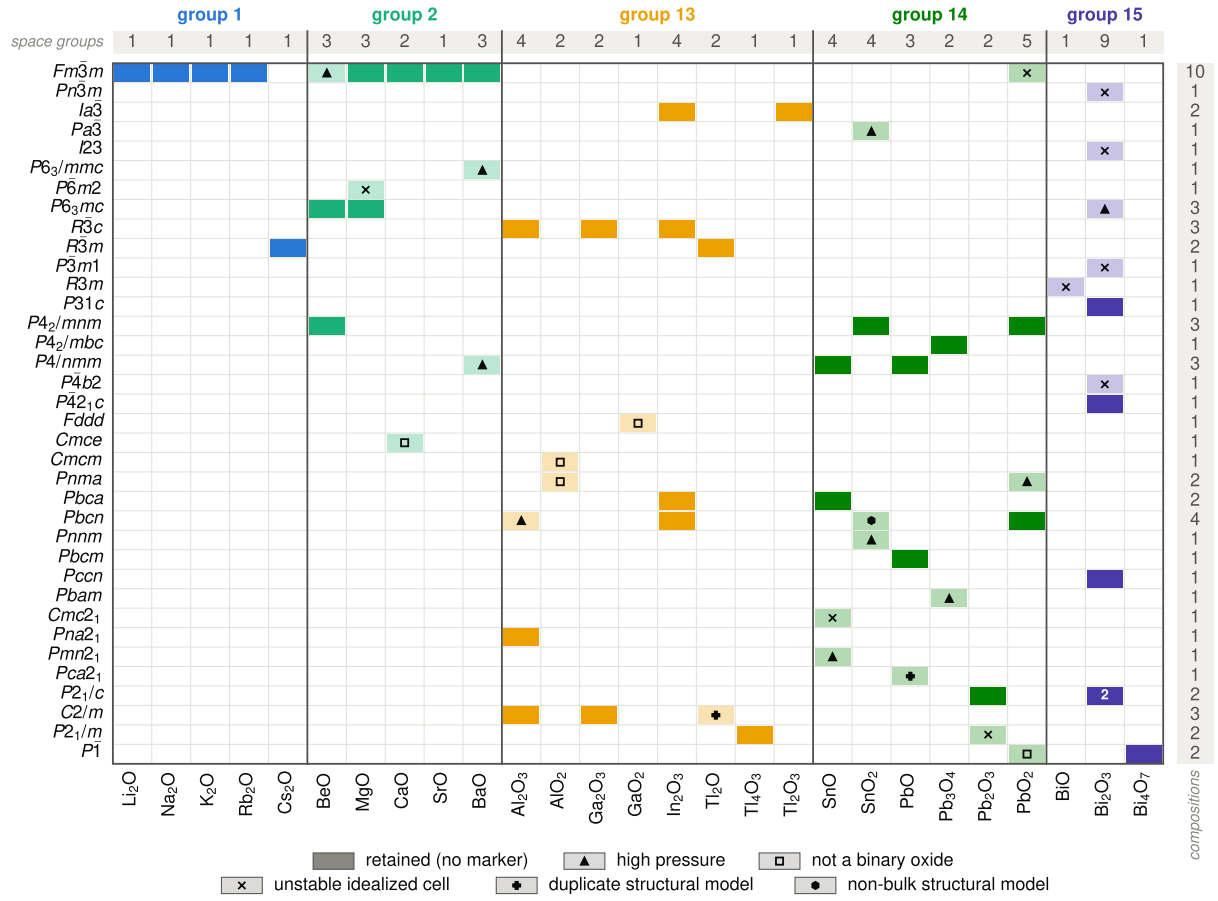


Figure 1: Composition–space-group coverage of the dataset and curation of the training set. Each filled cell holds one of the 66 material–polymorph combinations (the cell marked 2 holds two, α-Bi_2O_3 and Bi_2O_3-HP(I), both $P2_1/c$), colored by main group (groups 1, 2, 13, 14, and 15); compositions are ordered within each main group by cation atomic number and then by oxygen-to-metal ratio, and space groups are ordered by descending international number. Solid cells mark the 39 polymorphs retained for training; tinted cells mark the 27 excluded, with the marker giving the exclusion class as keyed in the legend. Marginal strips count space groups per composition (top) and compositions per space group (right).

Each pristine polymorph was relaxed, both ions and cell, in two independent runs started from the same Materials Project structure: the first from that structure unperturbed, and the second from a perturbed copy, in which both the atomic positions and the lattice vectors were displaced by Gaussian noise with a standard deviation equal to 10 % of the minimum interatomic distance. The forces on a symmetric configuration cannot break its symmetry, so only this "rattle" run can reach a lower-energy structure of different symmetry. It found one for ten of the 66 polymorphs, all of them excluded from training (SI Table S2). One of the two relaxed cells hosts the vacancy calculations of each retained polymorph and represents that polymorph everywhere below (cohesive energy, reduction potential, and convex hull); the chosen cell is the lower-energy of the two or lies within 1 meV atom^{-1} of it (SI Tables S1 and S2).

Vacancy Supercells, Configuration Search, and Formation Energies

Neutral oxygen vacancies were generated at every symmetry-inequivalent oxygen site of each relaxed polymorph with the PYMATGEN defect tools.[25,26] Each vacancy was modeled by removing a single oxygen atom from a supercell chosen as the most nearly cubic expansion of the relaxed cell, with 80 to 240 atoms (SI Table S1), and the atomic positions were relaxed at fixed cell shape and volume.

To prevent the defect from becoming trapped in a metastable local minimum, the local configuration space around each vacancy was explored with SHAKENBREAK using its default protocol.[16,17] The bonds between the vacancy and its two nearest cation neighbors were distorted by −60 % to 60 % of their equilibrium length in 10 % increments (13 levels per site) with a superimposed random rattle of every atom in the supercell, and each resulting configuration was relaxed independently, with the Brillouin zone sampled at the Γ point only; the sites that lack the energy of a level are identified in the Supporting Information. The lowest-energy relaxed configuration was retained as the ground-state vacancy structure.

This ground-state configuration and a pristine supercell of identical size and shape were then re-relaxed at fixed cell on a Γ-centered mesh, the production mesh (`KSPACING` = 0.22 Å^{-1}, the default spacing of the Materials Project r^{2}SCAN input set). All formation energies entering the fit are evaluated from this dense-matched pair of calculations. The pristine supercell of every polymorph in the fit was also re-relaxed at fixed cell from a rattled start, to test for a lower-energy configuration that a supercell tiled from the relaxed primitive cell cannot reach; none of the 39 references relaxed to a lower-energy structure (SI Table S3).

The neutral oxygen vacancy formation energy was evaluated as

$$E_\mathrm{v} = E_\mathrm{def} - E_\mathrm{bulk} + \mu_\mathrm{O}, \tag{1}$$

where E_def is the total energy of the relaxed defective supercell and E_bulk is the total energy of a defect-free supercell of identical size, shape, and k-point sampling.[27] The oxygen chemical potential μ_O is taken at the oxygen-rich limit, as one-half the total energy of an isolated, spin-polarized O_2 molecule computed at the same level of theory, $\mu_\mathrm{O} = \frac{1}{2}E(\mathrm{O_2}) = -5.930\,\mathrm{eV}$, with no empirical O_2 correction. Because r^{2}SCAN slightly overbinds O_2, omitting a correction shifts all *absolute* formation energies by a uniform ∼0.11 eV per oxygen; this constant is absorbed into the model intercept and affects neither the relative trends nor the fitted descriptor coefficients. No finite-size correction was applied to these neutral defects.[27]

Crystal Feature Model: Descriptor Construction

The neutral oxygen vacancy formation energy is predicted with the crystal feature model of Wexler et al.,[2] an interpretable linear regression whose four descriptors are the successive terms of the extended Born–Haber cycle introduced above: breaking the metal–oxygen bonds that coordinate the vacant site, accommodating the two liberated electrons by reducing neighboring cations or by populating the conduction band, and modulating both by the (in)stability of the host oxide. Two descriptors are *local* (evaluated from the immediate coordination environment of the vacant oxygen), and two are *global* properties of the host crystal.

The *crystal bond dissociation energy sum*, ΣE_b, quantifies the cost of severing the O–M bonds at the vacancy,[2]

$$\Sigma E_\mathrm{b} = \mathrm{CN_O} \cdot E_\mathrm{b}[\mathrm{O\text{–}M], \qquad E_\mathrm{b}[O\text{–}M] = \frac{E_\mathrm{c}[\mathrm{MO}_{n/2}]}{N_\mathrm{b}}}, \tag{2}$$

where $E_\mathrm{c}[\mathrm{MO}_{n/2}]$ is the cohesive energy of the lowest-energy retained polymorph of the binary oxide $\mathrm{MO}_{n/2}$, N_b is the number of O–M bonds per formula unit, and the oxygen coordination number $\mathrm{CN_O}$ is the number of

first-nearest-neighbor cations of the vacant oxygen, obtained with the CrystalNN algorithm.[28] The cohesive energy was computed by two independent routes: an r^{2}SCAN-consistent route, from spin-polarized isolated atoms (SI Table S9), and an experimental route, from tabulated 298 K thermochemistry[29,30] (Supporting Information).

The *crystal reduction potential*, V_r, quantifies how readily the host's cations accept the two liberated electrons. For the reduction $\mathrm{M}^{n+} \rightarrow \mathrm{M}^{m+}$ ($n > m$), referenced to $\frac{1}{4}\,\mathrm{O_2(g)}$, the gas-phase reservoir that also receives the vacancy oxygen in Eq. 1,

$$V_\mathrm{r} = \frac{-E_\mathrm{r}}{(n-m)\,F}, \qquad E_\mathrm{r} = E[\mathrm{MO}_{m/2}] + \frac{n-m}{4}\,E(\mathrm{O_2}) - E[\mathrm{MO}_{n/2}], \tag{3}$$

with F the Faraday constant and each oxide energy that of its lowest-energy retained polymorph.[2] Each composition is assigned a single couple, its adjacent recorded one: n is the average cation oxidation state of the host composition, and m is the next-lower oxidation state among the single-valence binary oxides of the same metal in the dataset (e.g., $\mathrm{PbO_2}$→PbO), or $m = 0$, reduction to the elemental metal, when no lower oxide is recorded.

In the perovskite parent, V_r is the maximum over the cations neighboring the vacant oxygen;[2] in a single-cation oxide, every neighbor is the same metal, so that maximum reduces to one value per composition, and site-to-site variation enters the descriptor set only through the coordination number in ΣE_b. As with E_b, V_r was evaluated on both the r^{2}SCAN and experimental energy scales; the two global descriptors below are r^{2}SCAN quantities in both routes.

The *band gap* E_g captures the cost of promoting electrons into the conduction band when they cannot be fully accommodated by cation reduction. We used the r^{2}SCAN fundamental gap, converged for every polymorph in the fit with a static calculation on a dense Γ-centered mesh (k-point spacing 0.11 Å^{-1}). Three hosts, $Ia\bar{3}$ $\mathrm{Tl_2O_3}$, rutile ($P4_2/mnm$) $\mathrm{PbO_2}$, and $Pbcn$ $\mathrm{PbO_2}$, are metallic within r^{2}SCAN and were assigned $E_\mathrm{g} = 0$.

The *energy above the convex hull* E_hull (in eV atom^{-1}) encodes the thermodynamic (in)stability of the host, a modulating factor first incorporated in the perovskite CFM.[2] It is computed at the r^{2}SCAN level, consistent with the vacancy energetics, from our own relaxed retained polymorphs together with the elemental-metal and $\mathrm{O_2}$ references. The current Materials Project hull energies[13] agree with these values to within 6 meV atom^{-1} for the room-temperature-stable hosts (SI Figure S12), and substituting them for E_hull leaves the cross-validated accuracy of both model variants unchanged, so database hulls can substitute for the descriptor when the model is applied to compositions outside our calculation set.

Crystal Feature Model: Regression and Validation

Mirroring the perovskite analysis,[2] two model variants are reported: an all-site model trained on all 72 sites, which we refer to as the clean sites, and a room-temperature-stable model restricted to the 56 sites whose host lies within 25 meV atom^{-1} ($\approx k_\mathrm{B}T$ at 298 K) of the convex hull. The four descriptors are combined in a linear model,

$$E_\mathrm{v} = c_\mathrm{b}\,\Sigma E_\mathrm{b} + c_\mathrm{r}\,V_\mathrm{r} + c_\mathrm{g}\,E_\mathrm{g} + c_\mathrm{hull}\,E_\mathrm{hull} + c_0, \tag{4}$$

fit with the Huber regressor,[31] a robust estimator whose loss is linear for large residuals so that outlying sites cannot dominate the fit, as implemented in SCIKIT-LEARN.[32] Following the perovskite study,[2] the model was cross-validated by averaging over 1000 independent fits, each trained on a random 50 % partition of the sites and evaluated on the held-out half; we report the mean and standard deviation of the held-out (test) mean absolute error (MAE) and of the coefficients, and models are compared on paired splits (the same 1000 partitions), by the mean change of the test MAE and the share of splits in which it decreases. On the present data, the bond-dissociation and hull coefficients of Eq. 4 are statistically null (see Results), so the primary model retains only V_r and E_g; the full four-descriptor form is reported alongside it, as the Born–Haber decomposition and for comparability with the perovskite parent.

Because three of the four descriptors take a single value per host, a random site-level partition places sites of the same host on both sides of the split. To confirm that this shared-host information does not inflate the held-out accuracy, leave-one-polymorph-out and leave-one-composition-out cross-validation was also performed, in which every site of a held-out group is predicted by a model trained without that group.

Pooled over all held-out sites, the resulting MAEs of the primary model, 0.30 eV (leave-one-polymorph-out) and 0.31 eV (leave-one-composition-out) for the room-temperature-stable model (0.28 and 0.29 eV for the all-site model), reproduce the random-split values, so the reported accuracy also describes prediction for compositions absent from the training set (SI Table S5).

Results and Discussion

Periodic Trends in Vacancy Formation Energies

The curated dataset yields 72 well-defined neutral oxygen vacancy formation energies across 24 binary main-group oxide compositions. Figure 2 organizes these energies by main group and composition, revealing trends down each main group and with cation oxidation state. The periodic trend is read from each composition's lowest-E_{hull} (ground-state) polymorph, the thermodynamically preferred phase, rather than from a coverage-weighted average over every computed site. The formation energies span 8 eV, from 0.09 eV for a vacancy in the rutile ($P4_2/mnm$) polymorph of PbO_2 to 8.09 eV in wurtzite BeO. Within each main group, the E_{v} of the ground-state polymorph decreases monotonically down the group as the cation grows and its charge density drops: from Li_2O (7.4 eV) to Cs_2O (3.8 eV) across group 1, and from BeO (8.1 eV) to BaO (5.9 eV) across group 2. Among the multivalent Tl, Pb, and Bi oxides, the ground-state-polymorph E_{v} also decreases monotonically with increasing cation oxidation state; Sn is the exception, with its two oxidation states within 0.3 eV of each other and the dioxide slightly higher. Metastable polymorphs of the same composition lie mostly below these ground-state values (open symbols in Figure 2) without disturbing the trend. The closed-shell, redox-inert oxides of groups 1 and 2 are the most vacancy-resistant (means over all sites of 5.2 eV for group 1 and 7.2 eV for group 2), whereas the reducible late-p-block oxides of groups 13–15 form vacancies far more readily; the group-14 (Pb, Sn) and group-15 (Bi) oxides are the lowest (all-site means of 2.7 eV and 2.8 eV, respectively), with the lowest sites (both PbO_2 polymorphs and Tl_2O_3) falling below 1 eV.

The scatter among the sites of a single polymorph in Figure 2 also reflects chemical differences among the groups. The group 1 and 2 oxides show no such scatter, because the isotropic, ionic bonding of these closed-shell cations favors high-symmetry structures (antifluorite, anti-$CdCl_2$, rock salt, wurtzite, and rutile-related) with a single symmetry-inequivalent oxygen site. The group 13 sesquioxides, whose smaller trivalent cations bond more covalently and directionally, instead adopt polymorph-rich, lower-symmetry structures whose inequivalent oxygen sites differ in local environment (within Al_2O_3 alone, oxygen coordination numbers of 3 to 5 and bond-dissociation sums ΣE_{b} spanning 8.1–13.4 eV), spreading E_{v} by as much as 0.9 eV within a single composition (Ga_2O_3). The site diversity that produces this scatter is the local-environment variation the crystal feature model's local descriptors are constructed to capture; within a single host, it is also the variation that they ultimately fail to capture (see the outlier analysis below).

The organizing variable is cation reducibility. Oxides whose metal can access a lower oxidation state ($Pb^{4+}\rightarrow Pb^{2+}$, $Tl^{3+}\rightarrow Tl^{+}$) form vacancies far more easily than oxides of the redox-inert closed-shell cations of groups 1 and 2, because the two electrons left behind by the departing oxygen can occupy an energetically accessible state. This is the qualitative signature that the crystal reduction potential V_{r} is constructed to quantify, and it motivates the descriptor analysis that follows.

The Crystal Feature Model

In isolation, three of the four descriptors correlate with E_{v} with the signs that the extended Born–Haber cycle predicts (SI Figure S10). The crystal reduction potential is the strongest single predictor (Pearson $R = -0.98$), reproducing its dominance in the perovskite CFM ($R = -0.88$).[2] The band gap is second and positive ($R = +0.88$), far stronger than in the perovskite study ($R = +0.34$),[2] because the main-group binaries span band gaps from zero (the three metallic hosts) to 8.5 eV, a far wider range than that of the transition-metal perovskites. The bond-dissociation sum correlates moderately ($R = +0.76$), whereas the energy above the hull correlates only weakly and with the opposite sign to the one the cycle predicts ($R = +0.30$).

Combining the four descriptors in the linear model of Eq. 4 reproduces the r^{2}SCAN formation energies, but two of the terms carry no statistical weight. For the room-temperature-stable hosts, the four-descriptor fit gives a held-out test MAE of 0.33(4) eV (mean over the 1000 splits; standard deviation in the last

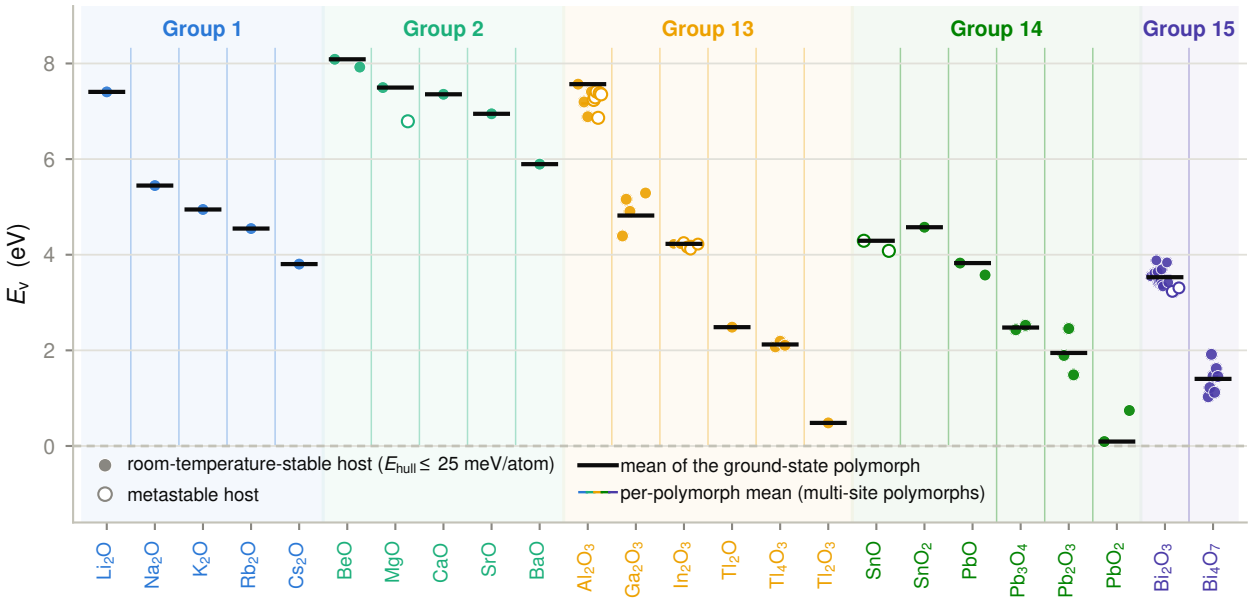


Figure 2: Neutral oxygen vacancy formation energies E_v across the binary main-group metal oxides. Each point is one symmetry-inequivalent oxygen site (72 clean sites across 39 polymorphs of 24 compositions). Compositions are blocked by main group and ordered within each group by cation atomic number, with same-cation stoichiometries ordered by increasing oxygen-to-metal ratio (increasing formal cation oxidation state). Within each composition, sites belonging to the same polymorph are clustered about a common horizontal offset, with polymorphs ordered left to right by increasing energy above the convex hull E_{hull} (the leftmost polymorph is the most stable); horizontal position otherwise carries no quantitative meaning. Short colored ticks give per-polymorph mean E_v for multi-site polymorphs in multi-polymorph compositions; the long black bar marks the mean E_v of each composition's lowest-E_{hull} (ground-state) polymorph, the value compared across compositions; for SnO, whose polymorphs all lie outside the room-temperature window, that polymorph is drawn with open symbols. The full polymorph list, with the number of sites per polymorph, is given in SI Table S1.

digit; $n = 56$), with bond-dissociation and hull coefficients both within two standard deviations of zero (the consensus fit to all those sites, an in-sample measure, has $R^2 = 0.97$). Trained on all 72 clean sites, metastable hosts included, it gives 0.30(3) eV (in-sample $R^2 = 0.97$) with the same two null terms. Restricting the model to its two surviving descriptors lowers the held-out error: on paired splits the test MAE changes by -0.014 eV for the room-temperature-stable sites (lower in 67 % of the splits) and by -0.011 eV over all sites (64 %). The primary model, fit to the room-temperature-stable sites, is

$$E_\mathrm{v} = -1.74\,V_\mathrm{r} + 0.25\,E_\mathrm{g} + 0.91, \tag{5}$$

with E_v and E_g in eV and V_r in V; it achieves a test MAE of 0.31(4) eV on those sites (in-sample $R^2 = 0.97$; Figure 3a); refit on all 72 clean sites, the same two-descriptor form gives 0.28(3) eV (in-sample $R^2 = 0.97$). The accuracies of the two training windows are statistically indistinguishable. The reported accuracy therefore does not depend on the stability restriction. Both values are lower than those of the four-descriptor transition-metal perovskite parent, which achieved 0.45(4) eV at the SCAN + U level for its room-temperature-stable set and 0.70(4) eV over all perovskites,[2] although the two settings differ: the present dataset contains no transition metals, and the model has no empirical Hubbard correction.

The coefficients of the full four-descriptor fit recover the physical hierarchy of the extended Born–Haber cycle and justify the two-descriptor simplification (Figure 3b). The reduction potential carries by far the largest standardized weight (-1.60 eV per standard deviation of V_r, $\sigma_\mathrm{r} = 0.96$ V over the 56 room-temperature-stable sites; native $c_\mathrm{r} = -1.67$ in the room-temperature-stable fit), confirming that electron accommodation by cation reduction dominates the energetics, and its native coefficient has the same sign as the perovskite value (≈ -1.5) and a magnitude about 0.2 larger.[2] The band gap is the second most important term ($c_\mathrm{g} = +0.23$), on a meta-GGA scale of the same order as the SCAN + U-level $c_\mathrm{g} \approx 0.4$ of the parent model.

The bond-dissociation term is null ($c_\mathrm{b} = +0.049 \pm 0.040$): although ΣE_b correlates moderately with E_v on its own, V_r and E_g already capture its variance, and removing it changes the room-temperature-stable test MAE by -0.006 eV on paired splits (lower in 52 % of them). This parallels the perovskite study, whose simplified room-temperature-stable model (built for the case of unknown E_g and retaining only V_r and E_hull) dropped the bond term because its coefficient, -0.03 ± 0.15, was statistically indistinguishable from zero. The hull coefficient is also null: its mean lies within two standard deviations of zero in both the room-temperature-stable fit ($+6.2 \pm 6.0$) and the all-site fit ($+2.9 \pm 4.3$), and it is as roster-dependent here as in the perovskite model, where it ranged from -5.8 to -55.8.[2] This instability reflects redundancy rather than missing physics: a signed counterpart of E_hull, the per-oxygen energy of reduction to the adjacent oxygen-poorer phases on our r^{2}SCAN convex hull, is collinear with V_r ($R = -0.999$ over all 72 clean sites; Supporting Information). E_hull therefore contributes only the polymorph-level metastability that V_r does not already encode; within the room-temperature window, that remainder spans 0–23 meV atom^{-1}, too narrow a range to constrain a coefficient.

Dropping both null terms yields the primary model of Eq. 5; restricting the fit to V_r alone, by contrast, degrades the test MAE to 0.37(5) eV (a paired change of $+0.055$ eV, lower in only 7 % of the splits), so the band gap contributes independent predictive information. Finally, models built from the experimental-thermochemistry route, which changes only ΣE_b and V_r, have test MAEs 0.04 to 0.07 eV above those of the r^{2}SCAN route in the four fits (two- and four-descriptor, both windows), so V_r can be sourced from tabulated thermochemistry at a modest cost in accuracy; the formation enthalpies of the mixed-valence Pb_2O_3, Tl_4O_3, and Bi_4O_7, which carry 13 of the sites, are estimates rather than tabulated values.

The reported accuracy was also verified to be robust to the methodological choices that most affect the target energies or the band-gap descriptor: the pristine reference, the band-gap definition, and the production k-point sampling (SI Figure S11). Evaluated on the four-descriptor form and on the same paired splits, the test MAE moves from 0.42(7) eV with a Γ-only perfect-supercell reference to 0.33(5) eV with the adopted convention. Varying the stability window shows the accuracy is not an artifact of the window choice: as the E_hull cutoff is relaxed from 25 meV atom^{-1} through 50 and 100 meV atom^{-1} to no cutoff at all, the within-window test MAE of the primary model takes the values 0.31, 0.28, 0.28, and 0.29 eV (Supporting Information). We retain the 25 meV atom^{-1} window for comparability with the parent model.

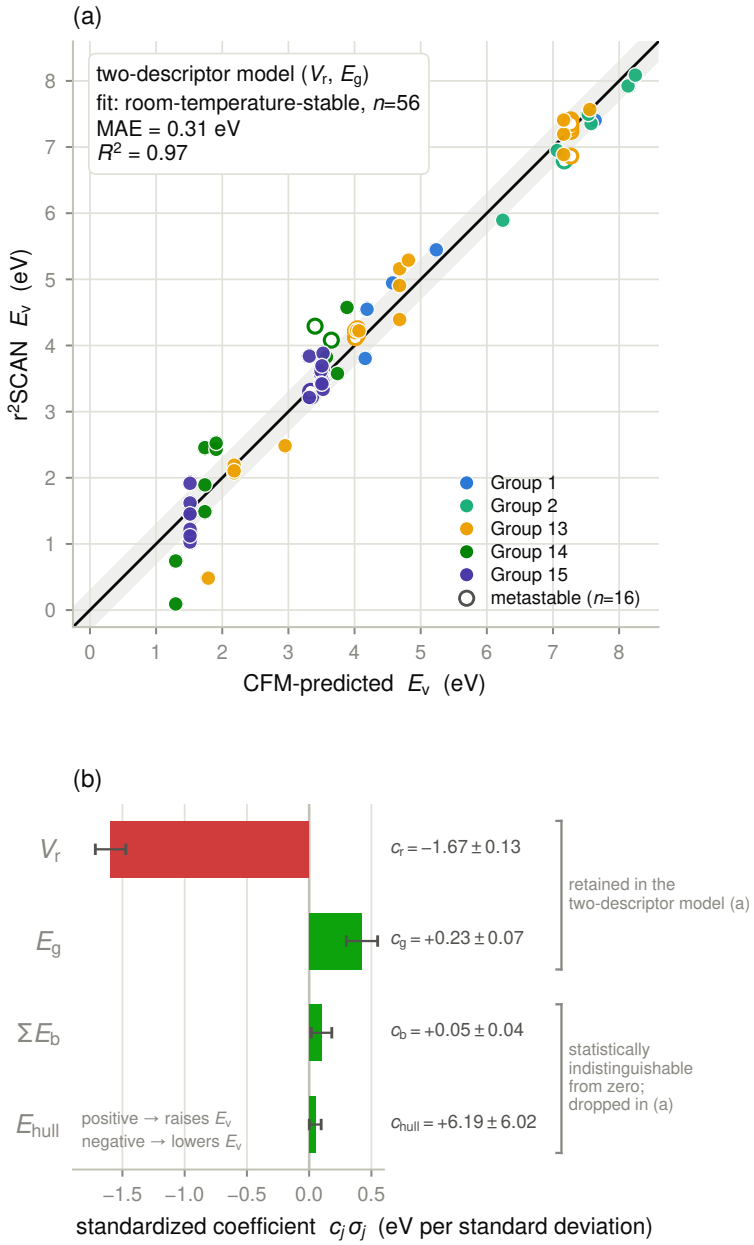


Figure 3: The binary-oxide crystal feature model (CFM). (a) r^{2}SCAN versus predicted E_v for the primary two-descriptor model of Eq. 5, colored by main group. Filled points are the 56 room-temperature-stable sites on which the primary model is fit (held-out test MAE 0.31 eV, $R^2 = 0.97$); open points are the 16 additional metastable sites, which lie outside the fit, so their predictions are out of sample (MAE 0.21 eV). The band marks ±MAE about the parity line. (b) Standardized coefficients $c_j\sigma_j$ of the full four-descriptor room-temperature-stable fit, each the change in E_v per one standard deviation σ_j of its descriptor (0.96 V for the reduction potential V_r, 1.85 eV for the band gap E_g, 7.9 meV atom^{-1} for the hull energy E_{hull}, and 2.04 eV for the bond-dissociation sum ΣE_b); the native-unit coefficients c_j and their fit uncertainties are annotated at the right.

Outliers, Within-Host Trends, and the Limits of a Linear Model

An interpretable model's residuals can be read mechanistically. Those of this model separate into three classes, each examined as a distinct route to extending the model: computational artifacts, which the residuals themselves exposed and curation removed or corrected; a systematic limit on how finely the model distinguishes sites within the same host, probed by searching for a fifth, site-resolved descriptor; and genuine electron-accommodation physics beyond the descriptor set, failures in *both* directions that together identify the missing term.

Computational Artifacts in the Residuals

Two vacancy calculations exposed unstable structural models of their hosts; in a fit that includes them, their three sites carry the largest residuals of the dataset. In the fluorite-type ($Fm\bar{3}m$) cell of PbO_2 and in the orthorhombic ($Cmc2_1$) cell of red SnO, the vacancy triggered a global reconstruction of the entire host rather than any local rebonding (SI Figures S8 and S9), and both cells are themselves unstable in our relaxations. Both polymorphs were therefore excluded during curation; their collapses are a caution that a vacancy energy computed in an incorrect or unstable structural model of the host reflects the relaxation of the host away from its assumed cell rather than defect chemistry. The *Pbca* cell of red SnO from the single-crystal redetermination, computed with the identical workflow and retained in the training set, shows none of this pathology.

Within-Host Degeneracy and the Search for a Site-Resolved Descriptor

The second class is subtler and systematic: the two-descriptor model is exactly flat *within* each polymorph, because both of its descriptors are host-level quantities (in the four-descriptor form, site-to-site discrimination appears only through ΣE_b, whose null coefficient suppresses it). The observed within-host spreads are substantial: the seven sites of Bi_4O_7 span 0.89 eV, the three sites of monoclinic Ga_2O_3 span 0.77 eV, and the three sites of the mixed-valence Pb_2O_3 span 0.97 eV, each set at a single predicted value, so this descriptor degeneracy sets a floor on the within-host resolution of any model built from these quantities. We therefore searched systematically for a fifth, site-resolved descriptor (SI Figure S13).

Each candidate was judged on host-centered values: the mean over a polymorph's sites is subtracted from both the descriptor and E_v, leaving only site-to-site differences (47 points from the 14 polymorphs with at least two sites). A single global coefficient can describe a within-host trend only if that trend has the same sign in every host, so each candidate's correlation was also counted host by host, over the hosts with at least three sites (the $+a/-b$ annotations of SI Figure S13a). The candidates are initial-state properties of the pristine crystal (the reduction potential of the vacancy's cation shell, coordination numbers and bond sums, bond lengths and bond-valence sums, the Madelung potential, the geometry of the cation shell, and the O-2p band center) and final-state properties of the relaxed defect supercell (the Kohn–Sham donor depth, the localization of the defect level, and the relaxation displacements); their definitions are given in the Supporting Information. Most candidates fail not because their correlations are weak but because their signs are inconsistent across hosts, so a global coefficient fit to one subset of hosts mispredicts the rest.

Among the sign-consistent candidates, the shortest metal–oxygen bond length of the vacant site, d_{min}, is the strongest: its host-centered $R = +0.65$ stays between +0.60 and +0.67 when any one host is left out, and its correlation is positive in each of the nine hosts with at least three sites (SI Figure S13). Within a host, the overbonded oxygen sites, those with the shortest bonds, are systematically the cheaper vacancies, a local strain-relief signature that runs opposite to the bond-breaking expectation encoded in ΣE_b. The bond length itself, not a proxy, carries the signal: within hosts, with the integer CN_O, $R(E_v, d_{min} \mid CN_O) = +0.57$ whereas $R(E_v, CN_O \mid d_{min}) = +0.05$, so the coordination number is a proxy for the bond length, not the reverse. This ordering does not, however, translate into accuracy: adding d_{min} to the primary model changes its held-out test MAE on the room-temperature-stable sites by only -0.024 eV, because the magnitude of the within-host trend differs from host to host (Supporting Information).

A site-resolved reduction potential makes the model worse. Replacing the composition-level V_r by an oxidation-state-resolved average over the vacancy's first cation shell (which differs from the composition value only in the mixed-valence hosts) *degrades* the four-descriptor model (test MAE 0.33 eV $\rightarrow$ 0.48 eV on the room-temperature-stable sites). The local and global trends need not share a sign: across compositions,

a more reducible cation lowers E_v, but in Pb_2O_3, where Pb^{4+} is by far the more reducible of the two cations, vacancies coordinated by more Pb^{4+} are *costlier* to form (SI Figure S7), because the stronger binding of oxygen to its higher-charged neighbors outweighs the greater reducibility of the shell. The host-level V_r is thus the appropriate formulation: the adjacent reduction couple available to the two released electrons is a property of the crystal as a whole, not an average over the vacancy's first coordination shell.

Outside the bond-geometry family, the Kohn–Sham donor depth is the only sign-consistent candidate, but it overlaps with d_{min}: in leave-one-host-out tests, the two together recover 44 % of the host-centered variance, against 39 % for d_{min} alone. No single descriptor from a published library[33] passes the false-discovery-rate, sign-consistency, and leave-one-host-out tests (Supporting Information).

The final-state electronic structure limits what any site-resolved descriptor can recover, which is why the search saturates, and at the same time confirms the Born–Haber picture the model is built on. The neutral vacancy forms no polaron at any site, and its defect level is a single doubly occupied band at 70 of the 72 sites, deep in the gap at all but four sites of the gapped hosts. At 22 of the 72 sites, the metal component of the level is predominantly *s*-like (more than half of the metal weight), the cation-reduction step of the cycle made directly visible. The site-to-site ordering of E_v thus tracks the depth of this defect level, which in turn tracks the local bond geometry that d_{min} encodes, but the *magnitude* of that dependence is host-specific. The remaining within-host variation (a root-mean-square value of 0.18 eV in the room-temperature window after a pooled d_{min} term fitted to these same sites) is therefore irreducible for any single global linear term and points instead to hierarchical, per-host-calibrated extensions.

The Electron-Accommodation Boundary

The third class carries the two largest residuals among the room-temperature-stable sites and identifies the physics the two descriptors approximate most coarsely: electron accommodation. The three hosts that r^{2}SCAN predicts to be metallic, $Ia\bar{3}$ Tl_2O_3, rutile ($P4_2/mnm$) PbO_2, and *Pbcn* PbO_2, have the three lowest formation energies in the dataset (0.48, 0.09, and 0.74 eV), which the primary model overpredicts by 0.6–1.3 eV (Figure 3a). These low energies reflect how the released electrons are accommodated. In each host, the two electrons released by the vacancy fill the metallic conduction states at the Fermi level rather than localizing on a reduced cation: the ShakeNBreak distortion sets are flat (every relaxed distortion lies within 6 meV of the ground state, apart from one +60 % Tl_2O_3 endpoint that relaxed into a distinct O–O-dimerized configuration 7.5 eV higher) and the local magnetic moments are zero. Independent defect studies corroborate both the low formation energy and the mechanism: hybrid-functional calculations find the oxygen vacancy in PbO_2 and Tl_2O_3 to be an always-ionized donor whose level is resonant in the conduction band (with a neutral formation energy of only 0.8 eV in Tl_2O_3 even under oxygen-rich conditions),[34,35] and experiment finds PbO_2 intrinsically oxygen-deficient: neutron diffraction resolves 1.6 % oxygen substoichiometry.[34] Facile vacancy formation with conduction-band electron accommodation is thus the established chemistry of these hosts; at the hybrid level, rutile PbO_2 (0.23 eV) and Tl_2O_3 (0.33 eV) are small-gap degenerate semiconductors rather than r^{2}SCAN metals,[34,35] a distinction that shifts E_v by tenths of an electronvolt but leaves the accommodation physics unchanged.

The linear model cannot reproduce such low values. Every reduction potential in the dataset is negative and $c_r < 0$, so the reduction term always *raises* the predicted energy; with $E_g = 0$ for these metallic hosts, the model cannot predict a formation energy below the intercept of Eq. 5 (0.91 eV), and all three computed metallic-host values lie below that floor, a limit built into the model's functional form rather than a failure of the descriptors.

Electron accommodation can also fail in the opposite direction. Litharge SnO ($P4/nmm$; a metastable host outside the room-temperature window) is the mirror image of the metallic hosts: the primary model *under*predicts it by 0.9 eV (predicted 3.40 versus a computed 4.29 eV), because its reducible Sn^{2+} couple and small band gap suggest low-cost electron accommodation that the actual electronic structure does not provide. The two electrons instead enter a filled, Sn-5s-derived state at the top of the valence band rather than executing the $Sn^{2+} \rightarrow Sn^0$ reduction assumed in V_r. An independent hybrid-functional study corroborates both the value and its interpretation: re-referenced to our oxygen chemical potential, the hybrid-functional neutral vacancy formation energy of Allen et al.[36] is ≈4.0 eV, within 0.3 eV of our r^{2}SCAN result, and the same study finds the oxygen vacancy in SnO to be a nondoping defect;[36] at the hybrid level, its donor levels lie below the valence-band maximum,[36] so the neutral vacancy we compute is also the physically relevant

charge state at every attainable Fermi level.

The two failure directions thus bracket a single deficiency, which lies in the accommodation route rather than in the band gap itself: V_r and E_g are host-level proxies for electron accommodation, and the true cost of accommodating the two liberated electrons can be lower than they encode (itinerant states at the Fermi level of a metallic host) or higher (a filled state at the top of the valence band). An explicit accommodation-energy term is therefore the extension of the descriptor set to which both failure directions point. Metallicity per se, however, is not the boundary of the linear model's validity: its perovskite parent was fit across hosts ranging from wide-band-gap insulators to metallic systems.[2] The boundary is the accommodation mode: the model remains valid wherever the liberated electrons localize on a cation redox couple that V_r describes, and fails where the accommodating state departs from that couple in either direction; conducting hosts whose carriers localize on a redox-active sublattice, as in many transition-metal oxides, should therefore remain within its reach.

Comparison with Prior Models

The CFM's accuracy of 0.31 eV is competitive with that of the random-forest models trained on much larger and more heterogeneous defect datasets (for example, the 0.27–0.44 eV of Kumagai et al.[9] across 937 oxides and the comparable accuracy of Baldassarri et al.[4] over more than 2500 vacancy calculations) but the CFM achieves it with two physically transparent descriptors, no defect calculation at prediction time, and residuals that can be read mechanistically. The Born–Haber form, transferred from transition-metal perovskites to transition-metal-free main-group binaries and re-evaluated at r^{2}SCAN, retains the coefficient hierarchy of its surviving terms while its bond-breaking and stability terms prove statistically null, so the four-descriptor parent reduces to a two-descriptor model with no loss of accuracy; the persistent dominance of the reduction potential supports the transferability of the underlying physical picture. This comparison is not like for like: the parent model predicted perovskite vacancies from the thermochemistry of the cations' binary oxides, a transfer between compound classes, whereas here the descriptors are bulk properties of the same binary oxides whose vacancies are predicted, and V_r in particular is built from the host's own bulk energy, so a lower error than the parent's is expected. The scope of the validation is set accordingly: the descriptors contain no defect calculation of any host, and the leave-one-composition-out error of 0.31 eV shows that a held-out composition's vacancy energetics are predicted from bulk properties alone. What the accuracy demonstrates is therefore the bulk-to-defect map within the binary-oxide class, not transfer between compound classes.

The descriptors themselves are not interchangeable with projected quantum-chemical bond-strength measures, which quantify only the covalent contribution to bond dissociation: on the 15 training-set hosts also covered by the LOBSTER bonding database of Naik et al.,[37] the integrated crystal orbital Hamilton population (ICOHP) summed over the bonds at the oxygen site is uncorrelated with both E_v ($R = +0.02$) and ΣE_b ($R = +0.05$; SI Table S8).

Conclusions

This work has surveyed neutral oxygen vacancy formation at the r^{2}SCAN level in 39 polymorphs of the binary main-group metal oxides and obtained 72 formation energies, ordered across the periodic table by the reducibility of the cation. A regression of these energies on the successive terms of the extended Born–Haber cycle assigns the bond-breaking and stability terms coefficients statistically indistinguishable from zero. The largest residuals among the room-temperature-stable sites arise where the liberated electrons, rather than reduce a cation, occupy itinerant conduction states in $Ia\bar{3}$ Tl_2O_3, rutile ($P4_2/mnm$) PbO_2, and $Pbcn$ PbO_2, whose formation energies are overpredicted. Electron accommodation fails in the opposite direction in litharge SnO, where a filled, Sn-5s-derived state at the top of the valence band leaves the formation energy underpredicted. The crystal feature model, linear in the crystal reduction potential and the band gap, has therefore proved valid wherever a cation redox couple accepts the liberated electrons. Within a host, the shortest metal–oxygen bond of the vacant site orders the formation energies in every polymorph with at least three sites, but the magnitude of that dependence is host-specific, so site-resolved prediction calls for per-host calibration rather than a fifth global descriptor.

Because its inputs are properties of the pristine bulk (and the reduction potential can be taken directly from tabulated experimental thermochemistry, which comes within 0.07 eV of the first-principles fit in test

MAE), the model supplies defect thermodynamics for compositions and polymorphs outside its training set at negligible cost, within the electron-accommodation boundary established above. Beyond the model, the dataset itself (72 curated, SHAKENBREAK-optimized vacancy formation energies spanning 8 eV of chemical variation) offers a benchmark for the machine-learning interatomic potentials now being extended to defect properties, where reference data for simple binary systems remain scarce; the model, in turn, provides a physics-based prior against which such potentials can be regularized or Δ-learned (trained to correct its residual) in that data-poor regime. Extension of both the dataset and the model to the transition-metal binary oxides, where multiple accessible oxidation states and localized d electrons will test the reduction-potential descriptor most severely, is underway and will be reported separately.

Data Availability Statement

The machine-readable data files described in the final section of the Supporting Information accompany this article as `cfm_si_data.zip`. The same tables, the analysis scripts that regenerate every table and figure, the figure files, and the software-environment specifications will be deposited in a Zenodo record on publication, under the CC BY 4.0 license for the data and figures and the MIT license for the scripts. The VASP input and output files of every production r^{2}SCAN calculation, except the `POTCAR` and wavefunction (`WAVECAR`) files, will be deposited in the NOMAD repository as a dataset on publication; the Supporting Information lists the calculations that it comprises.

Supporting Information

The Supporting Information is available free of charge.

Supporting Information (PDF): extended computational methods; the polymorph inventory with its ICSD provenance and the evidence for each curation outcome (SI Table S4); supercell dimensions and k-point sampling (SI Table S1); the space-group reproduction audit and the reference-supercell check (SI Tables S2 and S3); single-descriptor correlations and model-robustness tests (SI Figures S10 and S11 and Table S5); the comparison of r^{2}SCAN and Materials Project hull energies (SI Figure S12); the candidate site descriptors and their evaluation conventions (SI Tables S6 and S7); the comparison with quantum-chemical bonding descriptors (SI Table S8); and the isolated-atom reference energies (SI Table S9).

Machine-readable data files (ZIP): `cfm_si_data.zip`, holding the fitted coefficients and cross-validation statistics, the per-site predictions and descriptor values, the source data of the tables and figures, and the relaxed structures with their absolute total energies, each file described in the final section of the Supporting Information.

Acknowledgments

The authors thank the members of the Wexler group for helpful discussions. This research was supported by the U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, Hydrogen and Fuel Cell Technologies Office, under Award No. DE-EE0010733. A portion of the research was performed using computational resources sponsored by the Department of Energy's Office of Energy Efficiency and Renewable Energy and located at the National Renewable Energy Laboratory. The high-performance computing facilities at Washington University in St. Louis provided additional computational resources.

References

(1) Deml, A. M.; Stevanović, V.; Muhich, C. L.; Musgrave, C. B.; O'Hayre, R. Oxide Enthalpy of Formation and Band Gap Energy as Accurate Descriptors of Oxygen Vacancy Formation Energetics. *Energy Environ. Sci.* **2014**, *7*, 1996–2004.

(2) Wexler, R. B.; Sai Gautam, G.; Stechel, E. B.; Carter, E. A. Factors Governing Oxygen Vacancy Formation in Oxide Perovskites. *J. Am. Chem. Soc.* **2021**, *143*, 13212–13227.

(3) Emery, A. A.; Saal, J. E.; Kirklin, S.; Hegde, V. I.; Wolverton, C. High-Throughput Computational Screening of Perovskites for Thermochemical Water Splitting Applications. *Chem. Mater.* **2016**, *28*, 5621–5634.

(4) Baldassarri, B.; He, J.; Gopakumar, A.; Griesemer, S.; Salgado-Casanova, A. J. A.; Liu, T.-C.; Torrisi, S. B.; Wolverton, C. Oxygen Vacancy Formation Energy in Metal Oxides: High-Throughput Computational Studies and Machine-Learning Predictions. *Chem. Mater.* **2023**, *35*, 10619–10634.

(5) Linderälv, C.; Lindman, A.; Erhart, P. A Unifying Perspective on Oxygen Vacancies in Wide Band Gap Oxides. *J. Phys. Chem. Lett.* **2018**, *9*, 222–228.

(6) Tukur, P.; Wei, Y.; Zhang, Y.; Chen, H.; Lin, Y.; He, S.; Mo, Y.; Wei, J. Machine Learning-Guided Discovery of High-Entropy Perovskite Oxide Electrocatalysts via Oxygen Vacancy Engineering. *Small* **2025**, *21*, 2501946.

(7) Deml, A. M.; Holder, A. M.; O'Hayre, R. P.; Musgrave, C. B.; Stevanović, V. Intrinsic Material Properties Dictating Oxygen Vacancy Formation Energetics in Metal Oxides. *J. Phys. Chem. Lett.* **2015**, *6*, 1948–1953.

(8) Emery, A. A.; Wolverton, C. High-Throughput DFT Calculations of Formation Energy, Stability and Oxygen Vacancy Formation Energy of ABO_3 Perovskites. *Sci. Data* **2017**, *4*, 170153.

(9) Kumagai, Y.; Tsunoda, N.; Takahashi, A.; Oba, F. Insights into Oxygen Vacancies from High-Throughput First-Principles Calculations. *Phys. Rev. Mater.* **2021**, *5*, 123803.

(10) Witman, M. D.; Goyal, A.; Ogitsu, T.; McDaniel, A. H.; Lany, S. Defect Graph Neural Networks for Materials Discovery in High-Temperature Clean-Energy Applications. *Nat. Comput. Sci.* **2023**, *3*, 675–686.

(11) Choudhary, K.; Sumpter, B. G. Can a Deep-Learning Model Make Fast Predictions of Vacancy Formation in Diverse Materials? *AIP Adv.* **2023**, *13*, 095109.

(12) Mueller, T.; Montoya, J.; Ye, W.; Lei, X.; Hung, L.; Hummelshøj, J.; Puzon, M.; Martinez, D.; Fajardo, C.; Abela, R. An Electrochemical Series for Materials. *Proc. Natl. Acad. Sci. U. S. A.* **2024**, *121*, e2320134121.

(13) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, 011002.

(14) Zagorac, D.; Müller, H.; Ruehl, S.; Zagorac, J.; Rehme, S. Recent Developments in the Inorganic Crystal Structure Database: Theoretical Crystal Structure Data and Related Features. *J. Appl. Crystallogr.* **2019**, *52*, 918–925.

(15) Abbas, G.; Bulancea-Lindvall, O.; Davidsson, J.; Armiento, R.; Abrikosov, I. A. Theoretical Characterization of NV-like Defects in 4H-SiC Using ADAQ with SCAN and r^{2}SCAN Meta-GGA Functionals. *Appl. Phys. Lett.* **2025**, *126*, 154001.

(16) Mosquera-Lois, I.; Kavanagh, S. R.; Walsh, A.; Scanlon, D. O. ShakeNBreak: Navigating the Defect Configurational Landscape. *J. Open Source Softw.* **2022**, *7*, 4817.

(17) Mosquera-Lois, I.; Kavanagh, S. R.; Walsh, A.; Scanlon, D. O. Identifying the Ground State Structures of Point Defects in Solids. *npj Comput. Mater.* **2023**, *9*, 25.

(18) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for *Ab Initio* Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54*, 11169–11186.

(19) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50*, 17953–17979.

(20) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B* **1999**, *59*, 1758–1775.

(21) Furness, J. W.; Kaplan, A. D.; Ning, J.; Perdew, J. P.; Sun, J. Accurate and Numerically Efficient r^{2}SCAN Meta-Generalized Gradient Approximation. *J. Phys. Chem. Lett.* **2020**, *11*, 8208–8215.

(22) Furness, J. W.; Kaplan, A. D.; Ning, J.; Perdew, J. P.; Sun, J. Correction to "Accurate and Numerically Efficient r^{2}SCAN Meta-Generalized Gradient Approximation". *J. Phys. Chem. Lett.* **2020**, *11*, 9248.

(23) Kingsbury, R.; Gupta, A. S.; Bartel, C. J.; Munro, J. M.; Dwaraknath, S.; Horton, M.; Persson, K. A. Performance Comparison of r^{2}SCAN and SCAN metaGGA Density Functionals for Solid Materials via an Automated, High-Throughput Computational Workflow. *Phys. Rev. Mater.* **2022**, *6*, 013801.

(24) Kingsbury, R. S.; Rosen, A. S.; Gupta, A. S.; Munro, J. M.; Ong, S. P.; Jain, A.; Dwaraknath, S.; Horton, M. K.; Persson, K. A. A Flexible and Scalable Scheme for Mixing Computed Formation Energies from Different Levels of Theory. *npj Comput. Mater.* **2022**, *8*, 195.

(25) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Comput. Mater. Sci.* **2013**, *68*, 314–319.

(26) Shen, J.-X.; Voss, L. F.; Varley, J. B. Simulating Charged Defects at Database Scale. *J. Appl. Phys.* **2024**, *135*, 145102.

(27) Freysoldt, C.; Grabowski, B.; Hickel, T.; Neugebauer, J.; Kresse, G.; Janotti, A.; Van de Walle, C. G. First-Principles Calculations for Point Defects in Solids. *Rev. Mod. Phys.* **2014**, *86*, 253–305.

(28) Pan, H.; Ganose, A. M.; Horton, M.; Aykol, M.; Persson, K. A.; Zimmermann, N. E. R.; Jain, A. Benchmarking Coordination Number Prediction Algorithms on Inorganic Crystal Structures. *Inorg. Chem.* **2021**, *60*, 1590–1603.

(29) Barin, I. *Thermochemical Data of Pure Substances*, 3rd ed.; VCH: Weinheim, 1995.

(30) Chase, M. W., Jr. NIST-JANAF Thermochemical Tables, Table O-001: Oxygen, Monatomic (O). 1998; `https://janaf.nist.gov/tables/O-001.html`.

(31) Huber, P. J. Robust Estimation of a Location Parameter. *Ann. Math. Stat.* **1964**, *35*, 73–101.

(32) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

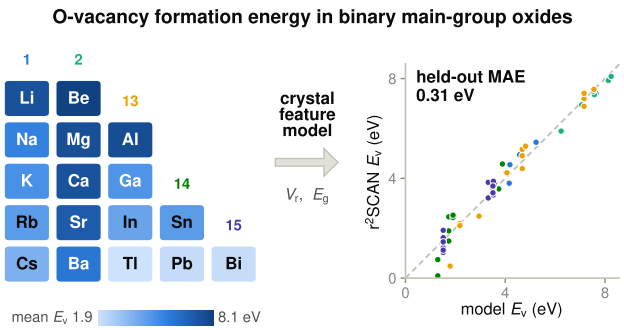
(33) Ward, L. et al. Matminer: An Open Source Toolkit for Materials Data Mining. *Comput. Mater. Sci.* **2018**, *152*, 60–69.

(34) Scanlon, D. O.; Kehoe, A. B.; Watson, G. W.; Jones, M. O.; David, W. I. F.; Payne, D. J.; Egdell, R. G.; Edwards, P. P.; Walsh, A. Nature of the Band Gap and Origin of the Conductivity of PbO_2 Revealed by Theory and Experiment. *Phys. Rev. Lett.* **2011**, *107*, 246402.

(35) Kehoe, A. B.; Scanlon, D. O.; Watson, G. W. Nature of the Band Gap of Tl_2O_3. *Phys. Rev. B* **2011**, *83*, 233202.

(36) Allen, J. P.; Scanlon, D. O.; Piper, L. F. J.; Watson, G. W. Understanding the Defect Chemistry of Tin Monoxide. *J. Mater. Chem. C* **2013**, *1*, 8194–8208.

(37) Naik, A. A.; Ertural, C.; Dhamrait, N.; Benner, P.; George, J. A Quantum-Chemical Bonding Database for Solid-State Materials. *Sci. Data* **2023**, *10*, 610.


O-vacancy formation energy in binary main-group oxides
1
2
13
14
15
Li
Be
Na
Mg
Al
K
Ca
Ga
Rb
Sr
In
Sn
Cs
Ba
Tl
Pb
Bi
mean E_v 1.9
8.1 eV
crystal feature model
V_f, E_g
held-out MAE 0.31 eV
r^{2}SCAN E_v (eV)
model E_v (eV)
0
4
8

Supporting Information: Periodic Trends in Oxygen Vacancy Formation Energies Across Binary Main-Group Metal Oxides

Manish Kumar[1], Bridget S. Ferris[1], Avital Isakov[1,2], and Robert B. Wexler*[1]

[1]Department of Chemistry and Institute of Materials Science and Engineering, Washington University in St. Louis, St. Louis, MO 63130, USA
[2]Present address: Department of Chemistry and Biochemistry, University of California, Los Angeles, Los Angeles, CA 90095, USA

*Email: wexler@wustl.edu

Extended computational methods

This section gives the computational methods in full; the Methods section of the main text is a condensed account of it. Figure S1 summarizes the workflow from source structure to fitted model.

Polymorph selection and training data curation

The dataset encompasses 66 material–polymorph combinations spanning 27 nominal binary oxide compositions of 17 main-group metals across groups 1 (Li, Na, K, Rb, Cs), 2 (Be, Mg, Ca, Sr, Ba), 13 (Al, Ga, In, Tl), 14 (Sn, Pb), and 15 (Bi) (main-text Figure 1). The polymorphs span all seven crystal systems and 36 space groups; $Fm\bar{3}m$, the most widely shared space group, appears in ten compositions across the alkali, alkaline-earth, and group-14 families. They include α-Bi_2O_3 (bismite), the ambient-pressure ground state of Bi_2O_3, which shares the space group $P2_1/c$ with the Bi_2O_3-HP(I) structure. A third $P2_1/c$ entry of Bi_2O_3 (mp-558953, 0.6 eV atom^{-1} above the Materials Project convex hull) was not computed, because it is not a binary oxide of O^{2-}: a third of its oxygen atoms form O–O pairs 1.52 Å apart, a peroxide-like separation. Because the crystal feature model (CFM)[S1] is best suited to hosts that are at least metastable, that is, at a local minimum of the energy under the near-ambient conditions of typical synthesis and operation, the training set was curated by systematically evaluating the experimental provenance of each polymorph using structural records from the Inorganic Crystal Structure Database (ICSD).[S2]

Experimental provenance

For each of the 66 material–polymorph combinations, every ICSD collection code that the Materials Project lists was retrieved, and the reported synthesis or measurement temperature and pressure were extracted where available (467 codes in total; 76 are computed structures, which were not counted, and 17 could not be retrieved from the ICSD; Table S4 enumerates every polymorph with its per-polymorph code count and curation outcome). Each polymorph was then classified according to whether at least one experimentally determined ICSD entry corresponded to near-ambient conditions, defined here as temperature within 273–313 K and pressure $\leq$0.2 MPa; computed structures were not counted, since the ICSD lists them at default ambient conditions when none are reported. Experimental entries without a recorded temperature or pressure carry the same ICSD default (293 K and 0.1 MPa) and were counted as near-ambient, and a phase quenched from high pressure was classified by the conditions of its recorded measurement; the Bi_2O_3 $P\bar{3}m1$ record, made at 6 GPa and measured at ambient pressure, is one such phase (see Excluded polymorphs). Polymorphs for which the ICSD records were restricted to near-ambient conditions or included a mixture of near-ambient and moderately elevated temperatures (up to $\sim$2200 K at ambient pressure) were retained for CFM training,

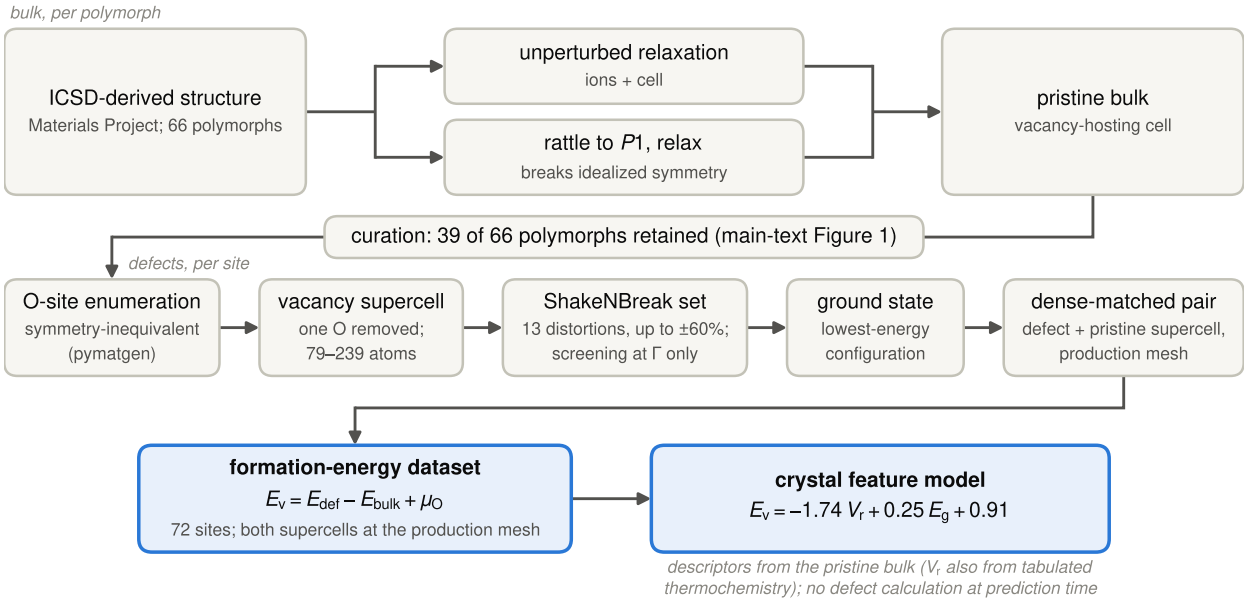


Figure S1: Computational workflow from ICSD-derived structure to fitted crystal feature model. Each polymorph's structure is relaxed twice from the same starting cell, once unperturbed and once after a symmetry-breaking rattle, and the relaxed cell of the stage that hosts the vacancy set serves as the pristine bulk (top row); the curation box, temperature–pressure provenance together with the structural checks of Polymorph selection, retains 39 of the 66 polymorphs for training. For every symmetry-inequivalent oxygen site, the ShakeNBreak ground-state configuration and a pristine supercell of identical size and shape are re-relaxed at the dense production mesh, forming the dense-matched pair (middle row). Every formation energy E_v entering the fit is evaluated from its dense-matched pair, and these energies train the crystal feature model (bottom row); the printed model equation, with V_r the crystal reduction potential and E_g the band gap, is the fit to the room-temperature-stable sites, main-text Eq. 5. Blue outlines mark the two products of the workflow: the formation-energy dataset and the fitted model.

as these represent thermodynamically accessible phases whose bonding environments are representative of the conditions under which vacancy formation energies are most commonly relevant. Where several ICSD entries exist for a polymorph, they were not compared structurally beyond their recorded conditions; each polymorph enters the calculations through its Materials Project structure, relaxed as described under Density functional theory calculations.

Excluded polymorphs

Of the 66 polymorphs evaluated, 39 were retained for CFM training, and 27 were excluded, in five categories (main-text Figure 1): (1) phases that exist only under high pressure, (2) records that are not binary oxides, (3) unstable idealized cells, (4) earlier structural models of retained phases, and (5) a model of a non-bulk layer. First, ten polymorphs were excluded because each exists exclusively under high hydrostatic pressure. These high-pressure phases, with the pressures of their ICSD structure determinations (for rock-salt BeO, the transition onset), are: BeO $Fm\bar{3}m$ (rock salt; wurtzite-to-rock-salt transition onset 137 GPa),[S3] BaO $P4/nmm$ (18.0–18.8 GPa),[S4,S5] BaO $P6_3/mmc$ (13.9 GPa),[S5] Al_2O_3 $Pbcn$ (113 GPa),[S6] Pb_3O_4 $Pbam$ (13.3 GPa; this record is phase III, with halved c, stable above about 6 GPa, not the $Pbam$ phase that minium adopts below 270 K at ambient pressure),[S7] PbO_2 $Pnma$ (23.5–140 GPa),[S8] SnO $Pmn2_1$ (9.0 GPa),[S9] SnO_2 $Pa\bar{3}$ (42–49 GPa),[S10] SnO_2 $Pnnm$ (12.6–15.5 GPa),[S10] and Bi_2O_3 $P6_3mc$ (2.8–5.4 GPa).[S11]

These polymorphs adopt compressed lattice geometries and modified coordination environments that differ substantially from ambient-pressure phases of the same composition. Including them in the training set would introduce bonding environments unrepresentative of the conditions under which the CFM descriptors (crystal bond dissociation energy, crystal reduction potential, band gap, and energy above the convex hull) are most physically meaningful. The descriptors, like the formation energies, are evaluated at zero pressure, at which these phases are at best metastable (rock-salt BeO, below, is dynamically unstable); at the pressures where the phases are stable, the vacancy formation enthalpy acquires a pressure–volume term, the product of the pressure and the vacancy relaxation volume, that none of the descriptors represents.

Rock-salt BeO warrants comment, because its ICSD record nominally reports ambient conditions: that record traces to a 1922 powder X-ray diffraction study (by the Debye–Scherrer method)[S12] which by its own account could not distinguish the rock-salt from the zinc-blende lattice for BeO despite many attempts. The same year, McKeehan[S13] assigned BeO the wurtzite type, and in 1925 Aminoff[S14] and Zachariasen[S15] established its hexagonal zinc-oxide-type lattice and rejected the rock-salt assignment. At ambient pressure, BeO is wurtzite-structured, with no transition observed below 55 GPa by Raman spectroscopy or 100 GPa in shock data, as summarized by Alptekin and Durandurdu.[S16] Consistent with a pressure-stabilized phase, the idealized rock-salt cell is dynamically unstable at zero pressure in our relaxations: its symmetry-broken (rattled) copy, relaxed as described under Density functional theory calculations, transforms to zinc blende, the cubic polytype of the tetrahedral wurtzite structure, 490 meV atom^{-1} lower in energy.

Second, five polymorphs were excluded because their sole ICSD provenance is not that of a binary oxide: the formally dioxide labels AlO_2 ($Cmcm$, $Pnma$) and GaO_2 ($Fddd$) correspond to two polymorphs of the oxyhydroxide AlOOH, boehmite[S17] and diaspore,[S18] and to GaOOH,[S19] in which the hydrogen atoms are absent from the recorded structure, so that the hydroxyl oxygen is counted as a second oxide oxygen; CaO $Cmce$ derives from a calcium–cobalt oxyhydroxide misfit-layer composite (a $CaOH/CoO_2$ intergrowth);[S20] and PbO_2 $P\bar{1}$ derives from the lead oxysulfate $Pb_3O_2(SO_4)$.[S21] A well-defined neutral oxygen vacancy formation energy does not exist for these misassigned compositions.

Third, nine polymorphs were excluded as unstable idealized cells: their computed cells are dynamically unstable, in the primitive cell or in the vacancy supercell, or do not reproduce the recorded structure, so that no vacancy formation energy of a distinct phase can be defined in them. The three Bi_2O_3 polymorphs $Pn\bar{3}m$,[S22] $I23$,[S23] and $P\bar{4}b2$[S24] are idealized models of intrinsically oxygen-disordered or off-stoichiometric phases (the disordered δ-Bi_2O_3 fluorite family, of which the $Pn\bar{3}m$ cell is an ordered-vacancy idealization, and the sillenite-type γ-Bi_2O_3), and MgO $P\bar{6}m2$[S25] (a hexagonal idealization of a stacking-disordered Mg/Al oxide) relaxes to the rock-salt MgO phase already present in the set. BiO $R3m$[S26] is a thin-film phase identified by electron diffraction, whose discoverers could establish neither its thermodynamic stability nor its range of existence and describe the rhombohedral modification as unstable; its relaxed cell is dynamically unstable in our relaxations, the symmetry-broken relaxation ending 237 meV atom^{-1} below the full-symmetry one (Table S2). For these five, a single ordered oxygen vacancy in the computed cell, and hence its formation

energy, is not well-defined.

The sixth is the orthorhombic $Cmc2_1$ model of red SnO, solved and refined only from powder data,[S27,S28] which has been superseded by a single-crystal redetermination of the red polymorph in the centrosymmetric space group $Pbca$[S29] and is dynamically unstable in our relaxations, losing the $Cmc2_1$ symmetry on symmetry-broken relaxation; the $Pbca$ redetermination itself is retained in the training set in its place.

Three further polymorphs were excluded on the same grounds. The La_2O_3-type ($P\bar{3}m1$) record of Bi_2O_3 is a phase made at 6 GPa and 880 °C and indexed from the powder pattern of the recovered sample with the atomic positions assumed from La_2O_3, which its authors describe as fairly unstable at room temperature;[S30] its pristine supercell, re-relaxed from a rattled start, lies 1.16 eV (per 90 atoms) below the symmetric cell (Table S2, note d). The monoclinic ($P2_1/m$) cell of Pb_2O_3 does not reproduce its only record (ICSD 36243): the Materials Project input cell is 9.5 % larger per formula unit, and its symmetric relaxation leaves $P2_1/m$. The fluorite-type ($Fm\bar{3}m$) cell of PbO_2 is unstable in the vacancy supercell: both rattled re-relaxations of its 96-atom pristine supercell fall below the ideal cell, by 69 and 96 meV atom^{-1} (Table S2, note e; its vacancy is discussed in the main text under Computational Artifacts in the Residuals).

Fourth, two polymorphs were excluded as earlier structural models of phases retained under another space group. The $Pca2_1$ record of PbO is an early X-ray model of massicot, superseded by neutron refinements in $Pbcm$,[S31] and, compared at the cell of each retained polymorph used in the fit, our relaxed $Pca2_1$ cell lies within 0.014 Å of the retained $Pbcm$ cell (largest atomic displacement after superposition; root-mean-square 0.012 Å). The $C2/m$ and $R\bar{3}m$ records of Tl_2O are two structural models of a single phase from the same study,[S32] and the relaxed $C2/m$ cell lies within 0.031 Å of the $R\bar{3}m$ cell (root-mean-square 0.030 Å); $R\bar{3}m$ Tl_2O is retained.

Fifth, the orthorhombic ($Pbcn$) SnO_2 polymorph was excluded because its only record (ICSD 62199) is a model of a non-bulk layer: a structure determined by reflection high-energy electron diffraction for an interface layer on quartz, not the bulk α-PbO_2-type phase.

Retained training set

Defect supercell calculations were carried out for the 63 polymorphs other than the three oxyhydroxides; vacancy formation energies were assembled for the 54 of these whose ICSD records do not identify them as exclusively high-pressure phases (among them rock-salt BeO, whose record nominally reports ambient conditions), and only the 39 retained polymorphs are used to fit and validate the CFM.

A small number of *included* polymorphs have individual ICSD entries recorded outside the near-ambient window (e.g., Li_2O $Fm\bar{3}m$ at 873 K, CaO $Fm\bar{3}m$ at 10 K) or at elevated pressures (e.g., BeO $P6_3mc$ up to 5 GPa, SnO_2 $P4_2/mnm$ up to 10.8 GPa, and corundum Al_2O_3 $R\bar{3}c$, whose 75 entries include compressed cells up to 113 GPa). In each of these cases, near-ambient conditions account for the largest share of the experimental ICSD entries for that polymorph (a majority except for litharge PbO $P4/nmm$, 9 of 19), and the polymorph itself is well-established as the ambient-pressure ground state or a commonly observed metastable phase.

We further retain the rutile-related BeO $P4_2/mnm$ (β-beryllia, with tetrahedral Be),[S33] whose only structural record is at high temperature. Unlike a high-pressure phase, its relaxed cell keeps this structure when its symmetry is broken (its rattled copy relaxes back to $P4_2/mnm$, isoenergetic to within 0.1 meV atom^{-1}; Table S2), and it lies only 16 meV atom^{-1} above the wurtzite ground state. It is therefore a well-defined ambient-pressure polymorph whose bonding environment provides a valid vacancy-formation data point; its occurrence at high temperature reflects entropic stabilization rather than any structural instability.

The retained training set thus comprises 39 polymorphs spanning 24 unique compositions. Enumerated with the PYMATGEN vacancy generator on the relaxed unperturbed cell of each, these polymorphs present 72 symmetry-inequivalent oxygen sites, at each of which a vacancy formation energy was computed (Regression and validation subsection). The sites included in the model span twofold through eightfold oxygen coordination (Figure S2), in contrast to the fixed oxygen environments of the perovskite lattice, so the bond-dissociation descriptor defined below varies from site to site through the coordination number.

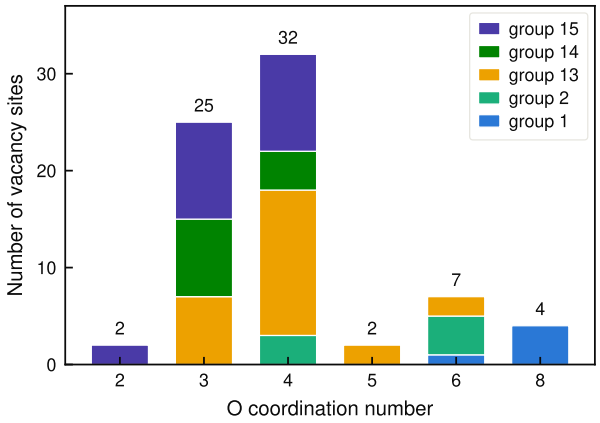


Figure S2: Local-environment diversity of the training set. Oxygen coordination numbers of the 72 clean vacancy sites (those of the all-site model; the primary model is fit to the 56 room-temperature-stable sites among them), stacked by main group with the colors of main-text Figure 1; totals are annotated above each bar. The coordination-number axis is categorical: only the coordination numbers that occur are shown, and no site is sevenfold coordinated.

Density functional theory calculations

All density functional theory (DFT) calculations were performed with the Vienna *Ab initio* Simulation Package (VASP, version 6.4; the builds used and their cross-check are described below, after the reference-supercell check of Table S3)[S34] using the projector augmented-wave (PAW) method.[S35,S36] Exchange and correlation were described by the r^{2}SCAN meta-generalized-gradient-approximation (meta-GGA) functional,[S37,S38] adopted for consistency with the r^{2}SCAN thermodynamic data of the Materials Project.[S39–S41] Input files were prepared with PYMATGEN[S42] following the Materials Project r^{2}SCAN relaxation set,[S40] which applies neither a Hubbard-U correction nor hybrid exchange. We employed the Materials Project's recommended PAW potentials, which are generated within the Perdew–Burke–Ernzerhof parametrization[S43] and treat semicore states as valence where appropriate: `Li_sv`, `Na_pv`, `K_sv`, `Rb_sv`, `Cs_sv`, `Be_sv`, `Mg_pv`, `Ca_sv`, `Sr_sv`, `Ba_sv`, `Al`, `Ga_d`, `In_d`, `Tl_d`, `Sn_d`, `Pb_d`, `Bi`, and `O`. The plane-wave basis was truncated at a kinetic-energy cutoff of 520 eV, and the aspherical gradient corrections inside the PAW spheres required for meta-GGA functionals were included (`LASPH` = `.TRUE.`). No symmetry was imposed in any calculation (`ISYM` = 0). The oxide, atom, and O_2 calculations were spin-polarized, started from the high-spin moments of the Materials Project input set (1 μ_B on each cation and 3 μ_B on each oxygen), and used Gaussian smearing with a width of 0.01 eV, with the Fermi level of insulators placed at midgap (`EFERMI` = `MIDGAP`). The electronic minimization used the all-band conjugate-gradient algorithm (`ALGO` = `All`) with at most 150 steps and projection in reciprocal space (`LREAL` = `.FALSE.`); the electronic self-consistency threshold was 1×10^{-5} eV, and geometries were optimized by a conjugate-gradient algorithm until all interatomic forces fell below 0.03 eV Å^{-1}. The band-gap calculations described below used a threshold of 1×10^{-6} eV; the few calculations that departed from these settings are listed below, after the reference-supercell check of Table S3, and the deposited input files record every calculation's settings. Where a variable-cell relaxation changed the volume within a single plane-wave basis, the energy of its final structure was recomputed in a fresh basis by a single-point calculation, and that energy is used throughout.

For each metal, the elemental reference is the lowest-energy phase among those computed: the Materials Project ground state and, for the alkali metals and tin, also the observed body-centered cubic and β-tin phases. Each phase was relaxed (ions and cell) with first-order Methfessel–Paxton smearing of 0.1 eV on a Γ-centered mesh with a k-point spacing of 0.10 Å^{-1}, and its energy was then computed without spin polarization by the tetrahedron method with Blöchl corrections on a mesh with a spacing of 0.050 Å^{-1} (0.075 Å^{-1} for Be and In); the energies on the two densest meshes computed for each phase agree to within 0.25 meV atom^{-1}. For tin, the lowest computed phase is α-tin: r^{2}SCAN places β-tin 126 meV atom^{-1} above it, against 15–

22 meV atom^{-1} from experiment and about 40 meV atom^{-1} in generalized-gradient calculations,[S44,S45] an overestimate of the α–β energy difference documented for r^{2}SCAN, which exceeds that of SCAN;[S45,S46] α-tin is the reference consistent with the 0 K oxide energies. A spin-polarized calculation started from 1 μ_B per atom lies at most 1.9 meV atom^{-1} below the non-spin-polarized energy (Rb). The metallic suboxides Cs_3O and Rb_9O_2, which enter only the convex hull, were treated in the same way.

For the bulk cells, Brillouin-zone integration used meshes generated automatically with PYMATGEN's density-by-reciprocal-volume scheme (`automatic_density_by_vol` with a grid density of 100 k-points per Å^{-3} of reciprocal-cell volume, the reciprocal lattice vectors including the factor 2π), which yields, for example, a Γ-centered $9 \times 9 \times 9$ mesh for cubic Li_2O. In the 2023 PYMATGEN release used to prepare the inputs, the scheme assigned Γ-centered meshes to 55 of the 66 polymorphs and Monkhorst–Pack meshes to the remaining 11 (7 of them among the 39 retained polymorphs; current releases give Na_2O a Γ-centered mesh). Each pristine polymorph was relaxed, both ions and cell, in two independent runs started from the same Materials Project structure, itself derived from the corresponding ICSD record.

The first run started from that structure unperturbed; since no symmetry was imposed, it is a symmetric start rather than a constraint, and its relaxed cell kept the starting space group for all but six polymorphs, all of them excluded from training (Table S2). Because meta-GGA relaxations seeded from idealized high-symmetry structures can stall at stationary points of the symmetric configuration, the second started from a perturbed copy, in which both the atomic positions and the lattice vectors were displaced by Gaussian noise with a standard deviation equal to 10 % of the minimum interatomic distance (of order 0.2 Å), lowering the symmetry to $P1$ and changing the starting volume by −11 to +16 % among the retained polymorphs. The forces on a symmetric configuration cannot break its symmetry, so only this "rattle" run can reach a lower-energy structure of different symmetry. It found one for ten of the 66 polymorphs, all of them excluded from training, and for four retained polymorphs it instead settled into a basin 39–378 meV atom^{-1} above the higher-symmetry cell (Table S2).

The vacancy formation energies of each polymorph that enter the fit come from one of its two relaxed cells, the higher-symmetry cell for 23 polymorphs and the rattled cell for 16 (Table S1), with the pristine reference of every defect calculation taken from the same cell, and that one cell represents the polymorph everywhere below (cohesive energy, reduction potential, and convex hull). It is the higher-symmetry cell wherever a vacancy set was computed in it and the rattled cell otherwise; the 43 vacancy energies computed in a fit polymorph's other cell are not used. Judged by the rule that the higher-symmetry cell is preferred unless the rattled one is more than 1 meV atom^{-1} lower in energy, the choice holds for every training polymorph: the chosen cell is the lower-energy of the two or lies within that margin of it (Table S2), and each rattled cell used relaxed back into the higher-symmetry structure (largest displacement 0.019 Å, volume within 0.5 %) except that of Cs_2O $R\bar{3}m$, a $C2/m$ shear of the layered stacking 0.45 meV atom^{-1} above the higher-symmetry cell, with a 1.9 % larger volume and a largest atomic displacement of 0.066 Å, which is used as nearly isoenergetic.

Vacancy supercells, configuration search, and formation energies

Neutral oxygen vacancies were generated at every symmetry-inequivalent oxygen site of each relaxed polymorph, with sites taken as equivalent within a symmetry tolerance of 0.01 Å, using the PYMATGEN defect tools.[S42,S47] Each vacancy was modeled by removing a single oxygen atom from a supercell that the PYMATGEN defect tools chose as the most nearly cubic expansion of the relaxed cell, not restricted to diagonal expansions, with 80 to 240 atoms (for Li_2O, a 240-atom $Li_{160}O_{80}$ supercell) and lattice vectors of at least 10 Å where these bounds allow; that length is reached in 35 of the 39 polymorphs that contribute sites to the fit, the shortest vector being 8.06 Å (BeO $P4_2/mnm$; the exceptions are listed in the next section). The atomic positions were then relaxed at fixed cell shape and volume. During this configuration screening, the Brillouin zone of every supercell was sampled at the Γ point only.

To prevent the defect from becoming trapped in a metastable local minimum, the local configuration space around each vacancy was explored with SHAKENBREAK using its default protocol.[S48,S49] The bonds between the vacancy and its two nearest cation neighbors (the default for a neutral oxygen vacancy) were distorted by −60 % to 60 % of their equilibrium length in 10 % increments with a superimposed random rattle of every atom in the supercell (standard deviation 10 % of the bulk nearest-neighbor distance, the SHAKENBREAK default), and each resulting configuration was relaxed independently. This gives 13 levels per site, including 0 %, which carries the same random rattle as the other levels, so that no unperturbed

configuration was relaxed; 24 of the 72 sites in the fit lack the energy of one or more levels of this protocol, 2 of them that of the +60 % level. The dimer configuration that later SHAKENBREAK versions add for vacancies was not generated, except at one site of *Pbca* SnO. At the most compressive levels (−60 and −50 %), two atoms of some starting structures were too close for VASP to run; at 19 sites the offending atoms were pushed away from their nearest neighbors until it ran, and three of the ground states in the fit come from such starts (the largest effect at site 8 of Bi_4O_7, 0.74 eV below the best unmodified configuration). The vacancy supercells and their distortion sets were generated with more than one version of the PYMATGEN defect tools and SHAKENBREAK, with the same distortion levels and rattle amplitude; the deposited input structures are those that were calculated. The lowest-energy relaxed configuration was retained as the ground-state vacancy structure. In prior high-throughput defect studies,[S50–S52] each vacancy was relaxed from a single starting configuration, either the unperturbed geometry or one with small random displacements of its neighbors, and Kumagai et al.[S50] re-relaxed from the symmetrized structure only when the symmetry increased.

This ground-state configuration and a pristine supercell of identical size and shape were then re-relaxed at fixed cell on a Γ-centered mesh, the production mesh, obtained by applying a uniform k-point spacing criterion to the supercell lattice (`KSPACING` = 0.22 Å^{-1}, the default spacing of the Materials Project r^{2}SCAN input set, which is at least as dense as any of the bulk-cell meshes above). All formation energies entering the fit are evaluated from this dense-matched pair of calculations; the per-polymorph supercell sizes and meshes are tabulated in Table S1. Before this final stage, the ground states of 22 of the sites had been re-relaxed on a coarser k-point mesh and the pristine references relaxed at the Γ point only; these intermediate references are the rungs of the reference comparison of Figure S11. The two cells of a pair share the k-point set but not always the spin state: in two pairs, the defect cell keeps a net moment against a nonmagnetic reference, BaO $Fm\bar{3}m$ site 8 (1.06 μ_B) and In_2O_3 *Pbcn* site 16 (0.11 μ_B). For BaO, the formation energy is that of the moment-bearing state, the lowest found: forcing a zero net moment raises the defect energy by 50 meV and a non-spin-polarized calculation by 0.40 eV.

The pristine supercell of every polymorph in the fit was also re-relaxed at fixed cell from a rattled start (the perturbation described above, without the lattice shear), to test for a lower-energy configuration that a supercell tiled from the relaxed primitive cell cannot reach. None of the 39 references relaxed to a lower-energy structure; the outcomes are given with the reference-supercell check below (Table S3).

The neutral oxygen vacancy formation energy was evaluated as

$$E_v = E_{def} - E_{bulk} + \mu_O, \tag{S1}$$

where E_{def} is the total energy of the relaxed defective supercell and E_{bulk} is the total energy of a defect-free supercell of identical size, shape, and k-point sampling.[S53] The oxygen chemical potential μ_O is taken at the oxygen-rich limit, as one-half the total energy of an isolated, spin-polarized O_2 molecule computed at the same level of theory, $\mu_O = \frac{1}{2}E(O_2) = -5.930$ eV, with no empirical O_2 correction. Because r^{2}SCAN slightly overbinds O_2 (computed electronic dissociation energy 5.43 eV versus the experimental 5.21 eV, obtained by adding the 0.098 eV zero-point energy[S54] back to the measured 5.116 eV[S55]), omitting a correction shifts all *absolute* formation energies by a uniform ∼0.11 eV per oxygen; this constant is absorbed into the model intercept and affects neither the relative trends nor the fitted descriptor coefficients. The neutral vacancies lie in the dilute, charge-neutral limit, which strongly reduces the electrostatic finite-size effects that dominate for charged defects, and no such correction was applied.[S53]

Crystal feature model: descriptor construction

The neutral oxygen vacancy formation energy is predicted with the crystal feature model of Wexler et al.,[S1] an interpretable linear regression whose four descriptors are the successive terms of the extended Born–Haber cycle introduced in the main text: breaking the metal–oxygen bonds that coordinate the vacant site, accommodating the two liberated electrons by reducing neighboring cations or by populating the conduction band, and modulating both by the (in)stability of the host oxide. Two descriptors are *local* (evaluated from the immediate coordination environment of the vacant oxygen), and two are *global* properties of the host crystal.

Local descriptors

The *crystal bond dissociation energy sum*, ΣE_b, quantifies the cost of severing the O–M bonds at the vacancy. Following the original definition,[S1] the energy of a single O–M bond is

$$E_\mathrm{b}[\mathrm{O–M}] = \frac{E_\mathrm{c}[\mathrm{MO}_{n/2}]}{N_\mathrm{b}}, \tag{S2}$$

where $E_\mathrm{c}[\mathrm{MO}_{n/2}]$ is the cohesive energy of the lowest-energy retained polymorph of the binary oxide $\mathrm{MO}_{n/2}$, each polymorph represented by the relaxed cell that hosts its vacancy set (the cell used everywhere for that polymorph), and N_b is the number of O–M bonds per formula unit, counted from the oxygen side as the sum of the CrystalNN coordination numbers of the oxygen atoms. Because every cation neighbor of the vacant oxygen in a single-cation binary oxide is the same metal,

$$\Sigma E_\mathrm{b} = \mathrm{CN_O} \cdot E_\mathrm{b}[\mathrm{O–M}], \tag{S3}$$

where the oxygen coordination number $\mathrm{CN_O}$ is the number of first-nearest-neighbor cations of the vacant oxygen. $\mathrm{CN_O}$ was obtained by applying the CrystalNN algorithm,[S56] as implemented in PYMATGEN,[S42] to a supercell tiled from the relaxed bulk cell. The integer neighbor count is a deliberate simplification; fractional, CrystalNN-weighted coordination numbers and the bond-valence sum of the removed oxygen are among the site-resolved alternatives assessed in the search for a site-resolved descriptor below.

The cohesive energy was computed by two independent routes: an r^{2}SCAN-consistent route, $E_\mathrm{c} = \sum_i E[\mathrm{atom}_i(g)] - E[\mathrm{MO}_{n/2}(s)]$, from spin-polarized isolated atoms with the ground-state magnetic moments dictated by Hund's rules (Table S9); and an experimental route, $E_\mathrm{c} = -\Delta H_\mathrm{f}[\mathrm{MO}_{n/2}] + E_\mathrm{c}[\mathrm{M}] + (n/4)\, D(\mathrm{O_2})$, from tabulated thermochemistry (each value is listed with its source in the deposited file `experimental_thermo.json`; Additional supporting data). Every term of this route is a 298 K enthalpy: $\Delta H_\mathrm{f}[\mathrm{MO}_{n/2}]$ is the enthalpy of formation of the oxide and $E_\mathrm{c}[\mathrm{M}]$ the enthalpy of formation of the gaseous metal atom (the sublimation enthalpy of the metal), both from the tables of Barin,[S57] and $D(\mathrm{O_2})$ is the atomization enthalpy of $\mathrm{O_2}$ (5.165 eV, twice the enthalpy of formation of atomic oxygen)[S55] rather than the 0 K spectroscopic value. Those tables do not include the mixed-valence $\mathrm{Pb_2O_3}$, $\mathrm{Tl_4O_3}$, and $\mathrm{Bi_4O_7}$; their enthalpies of formation are estimates, identified as such in `experimental_thermo.json`.

The *crystal reduction potential*, V_r, quantifies how readily the host's cations accept the two liberated electrons. For the reduction $\mathrm{M}^{n+} \rightarrow \mathrm{M}^{m+}$ $(n > m)$, referenced to $\frac{1}{4}\,\mathrm{O_2(g)}$, the gas-phase reservoir that also receives the vacancy oxygen in Eq. S1, rather than to the standard hydrogen electrode,

$$V_\mathrm{r} = \frac{-E_\mathrm{r}}{(n-m)\,F}, \qquad E_\mathrm{r} = E[\mathrm{MO}_{m/2}] + \frac{n-m}{4}\,E(\mathrm{O_2}) - E[\mathrm{MO}_{n/2}], \tag{S4}$$

with F the Faraday constant and each oxide energy that of its lowest-energy retained polymorph, again in the cell that hosts that polymorph's vacancy set.[S1] Each composition is assigned a single couple, its adjacent recorded one: n is the average cation oxidation state of the host composition (a mean over two oxidation states in the four mixed-valence compositions $\mathrm{Tl_4O_3}$, $\mathrm{Pb_3O_4}$, $\mathrm{Pb_2O_3}$, and $\mathrm{Bi_4O_7}$), and m is the next-lower oxidation state among the single-valence binary oxides of the same metal in the dataset (oxide-to-oxide reductions for the multivalent cations Pb, Sn, Tl, and Bi, e.g., $\mathrm{PbO_2}$→PbO and $\mathrm{Pb_3O_4}$→PbO), or $m = 0$, reduction to the elemental metal, when no lower oxide is recorded. For every composition except $\mathrm{SnO_2}$ the adjacent couple is also the most favorable (maximum V_r) reduction to a single-valence oxide or to the metal; for $\mathrm{SnO_2}$, the adjacent couple $\mathrm{SnO_2}$→SnO replaces the direct reduction to the metal, which is more favorable by 0.04 V. Throughout the text and Supporting Information, V_r denotes this adjacent-couple value.

In the perovskite parent, V_r is the maximum over the cations neighboring the vacant oxygen;[S1] in a single-cation oxide, every neighbor is the same metal, so that maximum, and equally the neighbor average, reduces to one value per composition. V_r is thus local by construction but host-level in practice, and site-to-site variation enters the descriptor set only through the coordination number in ΣE_b (an oxidation-state-resolved shell average for the mixed-valence hosts is assessed in the search for a site-resolved descriptor below). As with E_b, V_r was evaluated on both the r^{2}SCAN and experimental energy scales; the two global descriptors below are r^{2}SCAN quantities in both routes.

Global descriptors

The *band gap* E_g captures the cost of promoting electrons into the conduction band when they cannot be fully accommodated by cation reduction. We used the r^{2}SCAN fundamental gap, the conduction-band minimum minus the valence-band maximum, whether the gap is direct or indirect. Because a gap read on the k-point mesh of a bulk relaxation is only an upper bound on the fundamental gap, the gap of every polymorph included in the fit was converged with a self-consistent r^{2}SCAN static calculation at the relaxed geometry on a Γ-centered mesh 2.0 to 4.0 times denser in each direction than the relaxation mesh (k-point spacing 0.11 Å^{-1}). Three hosts, $Ia\bar{3}$ Tl_2O_3, rutile ($P4_2/mnm$) PbO_2, and $Pbcn$ PbO_2, are metallic within r^{2}SCAN (their fundamental gap converges below 0.05 eV on a dense mesh) and were assigned $E_g = 0$; the small gaps that hybrid functionals give two of them are discussed in the outlier analysis of the main text.

The *energy above the convex hull* E_{hull} (in eV atom^{-1}) encodes the thermodynamic (in)stability of the host, a modulating factor first incorporated in the perovskite CFM.[S1] It is computed at the r^{2}SCAN level, consistent with the vacancy energetics: each metal–oxygen convex hull is constructed from our own relaxed retained binary-oxide polymorphs, each represented by the relaxed cell that hosts its vacancy set, together with the elemental-metal and O_2 references described above; three observed phases that host no vacancies (Bi_2O_4, Cs_3O, and Rb_9O_2) complete it, and excluded polymorphs do not enter it. The phases of the Materials Project hull that our set lacks are of two kinds: oxygen-richer phases of 8 alkali and alkaline-earth metals, mostly peroxides and superoxides, whose host compositions are vertices of our hull, so that these phases could change those hosts' E_{hull} only by destabilizing the vertices themselves; and Sn_5O_6, a computed phase with no experimental record, which lies 27 meV atom^{-1} above our hull, so that including it changes no hull energy.

The current Materials Project hull energies[S39] (the database's default hull, version 2026.09.28) agree with the r^{2}SCAN hull energies to within 6 meV atom^{-1} for the room-temperature-stable hosts (compared polymorph by polymorph, together with the database's r^{2}SCAN-only hull, in Figure S12), and substituting them for E_{hull} leaves the cross-validated accuracy of both model variants unchanged (four-descriptor test MAE 0.33 and 0.30 eV for the room-temperature-stable and all-site fits); drawing the room-temperature window on the database hull instead admits two further hosts, with the effect on the primary model given in the hull-comparison section below. Database hulls can therefore substitute for the descriptor when the model is applied to compositions outside our calculation set.

Crystal feature model: regression and validation

Vacancy formation energies were computed at every symmetry-inequivalent oxygen site of the vacancy-hosting cell of each polymorph (see Density functional theory calculations; the cell used for each polymorph is listed in Table S1), 72 sites across the 39 retained polymorphs. In a rattled cell, small residual distortions can make formerly equivalent oxygen sites inequivalent; oxygen sites that are equivalent when the cell is analyzed at a symmetry tolerance of 0.05 Å were counted once, keeping the lower-energy image (three such pairs, whose two images differ in formation energy by at most 3.2 meV), so that every site is symmetry-inequivalent in the unperturbed structure. Every one of the 72 sites has a well-defined E_v; we refer to them as the clean sites.

Mirroring the perovskite analysis,[S1] two model variants are reported: an all-site model trained on all 72 sites, and a room-temperature-stable model restricted to the 56 sites whose host lies within 25 meV atom^{-1} ($\approx k_BT$ at 298 K) of the convex hull. Each variant was fit with both the r^{2}SCAN-consistent and the experimental-thermochemistry descriptor routes defined above.

The four descriptors are combined in a linear model,

$$E_v = c_b\,\Sigma E_b + c_r\,V_r + c_g\,E_g + c_{hull}\,E_{hull} + c_0, \tag{S5}$$

fit with the Huber regressor,[S58] a robust estimator whose loss is quadratic for small residuals and linear for large ones so that outlying sites cannot dominate the fit, as implemented in SCIKIT-LEARN[S59] with its default parameters except for the iteration limit (2000, against a default of 100). Descriptors were standardized to zero mean and unit variance (population standard deviation) within each training fold, and the fitted coefficients were back-transformed to the native units of Eq. S5. Following the perovskite study,[S1] the model was cross-validated by averaging over 1000 independent fits, each trained on a random 50 % partition of the

sites and evaluated on the held-out half; we report the mean and standard deviation of the held-out (test) mean absolute error (MAE) and of the coefficients. Each fit is trained on $\lfloor n/2 \rfloor$ of the n sites; MAEs are quoted to two decimals with the split-to-split standard deviation in the last digit, and models are compared on paired splits (the same 1000 partitions), by the mean change of the test MAE and the share of splits in which it decreases. The separate analyses below draw their own split sequences; their MAEs for the same model agree within about 4 meV.

On the present data, the bond-dissociation and hull coefficients of Eq. S5 are statistically null, their means lying within two split-to-split standard deviations of zero (see the main-text Results), and removing both terms lowers the held-out error on paired splits. The primary model therefore retains only V_r and E_g and is fit and cross-validated by the identical protocol. The full four-descriptor form is reported alongside it, as the Born–Haber decomposition and for comparability with the perovskite parent.

Because three of the four descriptors take a single value per host, a random site-level partition places sites of the same host on both sides of the split; to confirm that this shared-host information does not inflate the held-out accuracy, deterministic leave-one-polymorph-out and leave-one-composition-out cross-validation was additionally performed, in which every site of a held-out group is predicted by a model trained without that group. Pooled over all held-out sites, the resulting MAEs of the primary model, 0.30 eV (leave-one-polymorph-out) and 0.31 eV (leave-one-composition-out) for the room-temperature-stable model (0.28 and 0.29 eV for the all-site model), reproduce the random-split values, so the reported accuracy also describes prediction for compositions absent from the training set; averaged over the held-out compositions instead of the sites, the leave-one-composition-out MAE is 0.38 eV (0.40 eV), because a few compositions with few sites carry larger errors (Table S5).

Supercell dimensions and Brillouin-zone sampling

The construction of the vacancy supercells and the Γ-point sampling of the ShakeNBreak configuration screening are described in the extended methods above. Every lattice vector of the pristine supercells is at least 10 Å long except in BeO $P4_2/mnm$, Al_2O_3 $Pna2_1$, In_2O_3 $Ia\bar{3}$, and Tl_2O_3 $Ia\bar{3}$, whose shortest lattice vector is 8.06–9.2 Å.

The formation energies do not depend on the screening sampling: the ground-state configuration of every vacancy site included in the fit and a pristine supercell of identical size and shape were re-relaxed at the common production mesh, so the sampling is common to both terms of the difference in Eq. S1. Table S1 reports, for every polymorph that contributes sites to the fit, the pristine supercell size and multiplicity, the number of symmetry-inequivalent oxygen vacancy sites included in the fit, and the production mesh together with its number of k-points (no symmetry was used, `ISYM` = 0, so only time-reversal pairs are combined).

Table S1: Pristine supercell size and Brillouin-zone sampling for each of the 39 polymorphs that contribute sites to the crystal-feature-model fit. Stage is the bulk relaxation (hs: higher-symmetry; rattle: symmetry-broken) on which the vacancy set and its pristine reference supercell were built. Sites is the number of symmetry-inequivalent oxygen vacancy sites included in the fit. N is the number of atoms in the pristine supercell (the defective cell contains $N - 1$), and Mult. (multiplicity) is the number of relaxed bulk cells the pristine supercell contains (the primitive cell for every polymorph except K_2O and Rb_2O, whose relaxed cell contains two primitive cells). The production mesh is the Γ-centered grid obtained by applying a uniform k-point spacing criterion (`KSPACING` = 0.22 $Å^{-1}$) to the supercell, at which the ground-state defect configuration of every site and its pristine reference supercell were re-relaxed and from which every formation energy entering the fit is evaluated; N_k is the number of k-points of the production mesh with no symmetry used (`ISYM` = 0), so that only time-reversal pairs are combined. Space group is the Materials Project label of the polymorph; its relaxed fit cell keeps that space group at the 0.1 Å tolerance of Table S2 except for Cs_2O $R\bar{3}m$ (rattled cell C2/m).

Group	Compound	Space group	Stage	Sites	N	Mult.	Production	N_k
1	Cs_2O	$R\bar{3}m$	rattle	1	96	32	$2 \times 3 \times 3$	10
1	K_2O	$Fm\bar{3}m$	rattle	1	108	18	$3 \times 3 \times 3$	14
1	Li_2O	$Fm\bar{3}m$	rattle	1	240	80	$3 \times 3 \times 3$	14
1	Na_2O	$Fm\bar{3}m$	rattle	1	192	64	$3 \times 3 \times 2$	10
1	Rb_2O	$Fm\bar{3}m$	rattle	1	108	18	$2 \times 3 \times 3$	10
2	BaO	$Fm\bar{3}m$	rattle	1	128	64	$3 \times 3 \times 2$	10
2	BeO	$P4_2/mnm$	rattle	1	96	12	$4 \times 4 \times 4$	36
2	BeO	$P6_3mc$	rattle	1	240	60	$3 \times 3 \times 3$	14
2	CaO	$Fm\bar{3}m$	rattle	1	96	48	$3 \times 3 \times 3$	14
2	MgO	$Fm\bar{3}m$	rattle	1	160	80	$3 \times 3 \times 3$	14
2	MgO	$P6_3mc$	rattle	1	128	32	$3 \times 3 \times 3$	14
2	SrO	$Fm\bar{3}m$	rattle	1	96	48	$3 \times 3 \times 3$	14
13	Al_2O_3	C2/m	hs	3	160	16	$3 \times 3 \times 3$	14
13	Al_2O_3	$Pna2_1$	hs	6	80	2	$3 \times 4 \times 4$	26
13	Al_2O_3	$R\bar{3}c$	rattle	1	180	18	$3 \times 3 \times 3$	14
13	Ga_2O_3	C2/m	rattle	3	160	16	$3 \times 3 \times 3$	14
13	Ga_2O_3	$R\bar{3}c$	rattle	1	180	18	$3 \times 3 \times 3$	14
13	In_2O_3	$Ia\bar{3}$	hs	1	160	4	$2 \times 2 \times 4$	12
13	In_2O_3	Pbca	hs	3	160	4	$3 \times 3 \times 2$	10
13	In_2O_3	Pbcn	rattle	2	160	8	$3 \times 3 \times 2$	10
13	In_2O_3	$R\bar{3}c$	hs	1	180	18	$2 \times 3 \times 3$	10
13	Tl_2O	$R\bar{3}m$	hs	1	96	16	$3 \times 3 \times 3$	14
13	Tl_2O_3	$Ia\bar{3}$	hs	1	160	4	$2 \times 2 \times 4$	12
13	Tl_4O_3	$P2_1/m$	hs	3	84	6	$3 \times 2 \times 3$	10

continued on next page

Table S1 – continued

Group	Compound	Space group	Stage	Sites	N	Mult.	Production	N_k
14	Pb_2O_3	$P2_1/c$	hs	3	160	8	$3 \times 2 \times 2$	8
14	Pb_3O_4	$P4_2/mbc$	hs	2	224	8	$3 \times 2 \times 2$	8
14	PbO	P4/nmm	hs	1	108	27	$3 \times 3 \times 2$	10
14	PbO	Pbcm	hs	1	96	12	$3 \times 3 \times 3$	14
14	PbO_2	$P4_2/mnm$	hs	1	216	36	$3 \times 2 \times 2$	8
14	PbO_2	Pbcn	hs	1	96	8	$3 \times 3 \times 3$	14
14	SnO	P4/nmm	hs	1	108	27	$3 \times 3 \times 2$	10
14	SnO	Pbca	hs	1	96	6	$2 \times 3 \times 3$	10
14	SnO_2	$P4_2/mnm$	hs	1	216	36	$3 \times 2 \times 2$	8
15	Bi_2O_3	$P\bar{4}2_1c$	hs	2	160	8	$3 \times 2 \times 2$	8
15	Bi_2O_3	$P2_1/c$	hs	6	160	4	$2 \times 2 \times 3$	8
15	α-Bi_2O_3	$P2_1/c$	hs	3	160	8	$3 \times 3 \times 2$	10
15	Bi_2O_3	P31c	hs	2	160	8	$3 \times 3 \times 3$	14
15	Bi_2O_3	Pccn	hs	2	80	4	$3 \times 3 \times 3$	14
15	Bi_4O_7	$P\bar{1}$	hs	7	176	8	$3 \times 3 \times 2$	10

Space-group reproduction under symmetry-breaking relaxation

The two independent relaxations of each bulk cell (extended methods above) are labeled here *higher-symmetry* (hs), for the run started from the unperturbed Materials Project structure, and *rattle*, for the run started from the perturbed copy. Table S2 compares the Materials Project space-group label of each polymorph with the space groups of the r^{2}SCAN-relaxed hs and rattle cells (determined with SPGLIB at a symmetry tolerance of 0.1 Å and an angle tolerance of 5°; each stage's cell is the converged run that hosts its vacancy set or, failing that, the last converged run of its restart chain), along with the rattle-minus-hs energy difference. The outcome is assigned from that difference: a rattle cell that retains the observed space group, or changes label at $|\Delta E| < 1\,\mathrm{meV\,atom^{-1}}$, counts as reproduced; a lower-energy cell in a different space group ($\Delta E < -1\,\mathrm{meV\,atom^{-1}}$) as a symmetry change; and a higher-energy one ($\Delta E > 1\,\mathrm{meV\,atom^{-1}}$) as a rattle artifact.

Across all 66 polymorphs in the dataset, 49 reproduce the observed space group or a tolerance-relabeled equivalent of it, 10 undergo a genuine energy-lowering symmetry change on rattling, and 7 rattle configurations converge to a higher-energy basin and are therefore treated as artifacts. For two of the retained oxides, Al_2O_3 $C2/m$ and Al_2O_3 $Pna2_1$, the rattle vacancy campaign was built on a bulk cell trapped in a higher-energy local minimum; for these the higher-symmetry set is used instead, consistent with the "hs" stage listed for them in Table S1.

Figure S3 resolves these outcomes by main group and by cation lone-pair activity. The 10 energy-lowering symmetry changes, all in excluded polymorphs, fall 5 among the 19 lone-pair-active hosts and 5 among the 47 others (26 % versus 11 %; two-sided Fisher exact test, $p = 0.14$), a difference that is not significant at this sample size; what does cluster among the lone-pair-active hosts is the tolerance relabel (8 of 19, against 4 of the 41 hosts without a lone pair). Lone-pair activity is assigned from the formal cation oxidation state (Tl^+, Pb^{2+}, Sn^{2+}, and Bi^{3+} active; the mixed-valence oxides partial), so BiO, with formal Bi^{2+} ($6s^26p^1$), is counted among the hosts without a lone pair; counted as active, it would give $p = 0.06$.

Table S2: Materials Project space-group label of each polymorph compared with the r^{2}SCAN-relaxed higher-symmetry (hs) and symmetry-broken (rattle) bulk cells, together with the rattle-minus-hs energy difference ΔE and the outcome of the space-group reproduction audit.

Group	Compound	MP label	hs	rattle	ΔE (meV atom^{-1})	Outcome
1	Cs_2O	$R\bar{3}m$	$R\bar{3}m$	C2/m	0	reproduced[a]
1	K_2O	$Fm\bar{3}m$	$Fm\bar{3}m$	$Fm\bar{3}m$	0	reproduced
1	Li_2O	$Fm\bar{3}m$	$Fm\bar{3}m$	$Fm\bar{3}m$	0	reproduced
1	Na_2O	$Fm\bar{3}m$	$Fm\bar{3}m$	$Fm\bar{3}m$	0	reproduced
1	Rb_2O	$Fm\bar{3}m$	$Fm\bar{3}m$	$Fm\bar{3}m$	0	reproduced
2	BaO	$Fm\bar{3}m$	$Fm\bar{3}m$	$Fm\bar{3}m$	0	reproduced
2	BaO	P4/nmm	P4/nmm	P4/nmm	0	reproduced
2	BaO	$P6_3/mmc$	$P6_3/mmc$	$P6_3/mmc$	0	reproduced
2	BeO	$Fm\bar{3}m$	$Fm\bar{3}m$	$F\bar{4}3m$	−490	symmetry change
2	BeO	$P4_2/mnm$	$P4_2/mnm$	$P4_2/mnm$	0	reproduced
2	BeO	$P6_3mc$	$P6_3mc$	$P6_3mc$	0	reproduced
2	CaO	Cmce	Cmce	$P\bar{1}$	0	reproduced[a]
2	CaO	$Fm\bar{3}m$	$Fm\bar{3}m$	$Fm\bar{3}m$	0	reproduced
2	MgO	$Fm\bar{3}m$	$Fm\bar{3}m$	$Fm\bar{3}m$	0	reproduced
2	MgO	$P\bar{6}m2$	$P\bar{6}m2$	$Fm\bar{3}m$	−370	symmetry change
2	MgO	$P6_3mc$	$P6_3mc$	$P6_3mc$	0	reproduced
2	SrO	$Fm\bar{3}m$	$Fm\bar{3}m$	$Fm\bar{3}m$	0	reproduced
13	Al_2O_3[c]	C2/m	C2/m	Cc	128	rattle artifact
13	Al_2O_3	Pbcn	Pbcn	Pbcn	0	reproduced
13	Al_2O_3[c]	$Pna2_1$	$Pna2_1$	P1	378	rattle artifact
13	Al_2O_3	$R\bar{3}c$	$R\bar{3}c$	$R\bar{3}c$	0	reproduced
13	AlO_2	Cmcm	Cmcm	C2/m	−239	symmetry change

continued on next page

Table S2 – continued

Group	Compound	MP label	hs	rattle	ΔE (meV atom^{-1})	Outcome
13	AlO_2	Pnma	Pnma	Pnma	0	reproduced
13	Ga_2O_3	C2/m	C2/m	C2/m	0	reproduced
13	Ga_2O_3	R$\bar{3}$c	R$\bar{3}$c	R$\bar{3}$c	0	reproduced
13	GaO_2[b]	Fddd	C2/c	P1	−6	symmetry change
13	In_2O_3	Ia$\bar{3}$	Ia$\bar{3}$	P1	280	rattle artifact
13	In_2O_3	Pbca	Pbca	$Pca2_1$	39	rattle artifact
13	In_2O_3	Pbcn	Pbcn	Pbcn	0	reproduced
13	In_2O_3	R$\bar{3}$c	R$\bar{3}$c	R$\bar{3}$c	0	reproduced
13	Tl_2O	C2/m	C2/m	P$\bar{1}$	0	reproduced[a]
13	Tl_2O	R$\bar{3}$m	R$\bar{3}$m	C2/m	1	reproduced[a]
13	Tl_2O_3	Ia$\bar{3}$	Ia$\bar{3}$	Ia$\bar{3}$	0	reproduced
13	Tl_4O_3	$P2_1/m$	$P2_1/m$	$P2_1/m$	0	reproduced
14	Pb_2O_3	$P2_1/c$	$P2_1/c$	$P2_1/c$	0	reproduced
14	Pb_2O_3[b]	$P2_1/m$	P4/nmm	P1	61	rattle artifact
14	Pb_3O_4	$P4_2/mbc$	$P4_2/mbc$	$P2_1/c$	0	reproduced[a]
14	Pb_3O_4	Pbam	Pbam	Pbam	0	reproduced
14	PbO	P4/nmm	P4/nmm	P4/nmm	0	reproduced
14	PbO	Pbcm	Pbcm	$P2_1$	0	reproduced[a]
14	PbO[b]	$Pca2_1$	Pbcm	Pbcm	0	reproduced[a]
14	PbO_2[e]	Fm$\bar{3}$m	Fm$\bar{3}$m	C2/m	0	reproduced[a]
14	PbO_2[b]	P$\bar{1}$	C2/c	P1	124	rattle artifact
14	PbO_2	$P4_2/mnm$	$P4_2/mnm$	$P4_2/mnm$	0	reproduced
14	PbO_2	Pbcn	Pbcn	Pbcn	0	reproduced
14	PbO_2	Pnma	Pnma	Pm	89	rattle artifact
14	SnO	$Cmc2_1$	$Cmc2_1$	Pnma	−8	symmetry change
14	SnO	P4/nmm	P4/nmm	P4/nmm	0	reproduced
14	SnO	Pbca	Pbca	$Pca2_1$	0	reproduced[a]
14	SnO	$Pmn2_1$	$Pmn2_1$	$Pmn2_1$	0	reproduced
14	SnO_2	$P4_2/mnm$	$P4_2/mnm$	$P4_2/mnm$	0	reproduced
14	SnO_2	Pa$\bar{3}$	Pa$\bar{3}$	Pa$\bar{3}$	0	reproduced
14	SnO_2	Pbcn	Pbcn	Pbcn	0	reproduced
14	SnO_2[b]	Pnnm	$P4_2/mnm$	$P4_2/mnm$	0	reproduced[a]
15	Bi_2O_3[b]	I23	I$\bar{4}$3m	P1	−150	symmetry change
15	Bi_2O_3[d]	P$\bar{3}$m1	P$\bar{3}$m1	C2	0	reproduced[a]
15	Bi_2O_3	P$\bar{4}2_1$c	P$\bar{4}2_1$c	$P2_12_12$	0	reproduced[a]
15	Bi_2O_3	P$\bar{4}$b2	P$\bar{4}$b2	P1	−205	symmetry change
15	Bi_2O_3	$P2_1/c$	$P2_1/c$	$P2_1/c$	0	reproduced
15	α-Bi_2O_3	$P2_1/c$	$P2_1/c$	$P2_1/c$	0	reproduced
15	Bi_2O_3	P31c	P31c	P31c	0	reproduced
15	Bi_2O_3	$P6_3mc$	$P6_3mc$	P31c	−13	symmetry change
15	Bi_2O_3	Pccn	Pccn	$P2_12_12$	0	reproduced[a]
15	Bi_2O_3	Pn$\bar{3}$m	Pn$\bar{3}$m	Cm	−329	symmetry change
15	Bi_4O_7	P$\bar{1}$	P$\bar{1}$	P$\bar{1}$	0	reproduced

15	BiO	R3m	R3m	Pm	−237	symmetry change

The hs and rattle space groups were determined with SPGLIB at a symmetry tolerance of 0.1 Å and an angle tolerance of 5°. ΔE is rounded to the nearest meV atom^{-1}, with outcomes assigned from the unrounded values; a negative value means the rattle cell is lower in energy. Outcome: reproduced, the rattle cell retains the Materials Project label; symmetry change or rattle artifact, a different space group with the rattle cell lower or higher in energy, respectively, by more than 1 meV atom^{-1} (an artifact is a relaxation trapped in a higher-energy basin). [a] A different space-group label at $|\Delta E| <$ 1 meV atom^{-1} (same structure; the tolerance-relabel class of Figure S3). [b] Even the hs relaxation departed from the Materials Project label at the 0.1 Å tolerance. [c] The rattle vacancy set is built on a bulk trapped in a higher-energy local minimum and is superseded by the hs set. [d] Reproduced in the primitive cell, but the 90-atom reference supercell relaxes 12.9 meV atom^{-1} lower from a rattled start (an off-centering of the interlayer oxygen atoms with a period longer than the primitive cell; see the discussion of the reference supercells that introduces Table S3); the polymorph is excluded as an unstable idealized cell. [e] Reproduced in the primitive cell, but both rattled re-relaxations of the 96-atom pristine supercell fall below the fluorite cell, by 69 and 96 meV atom^{-1}; the polymorph is excluded as an unstable idealized cell (main-text Polymorph Selection).

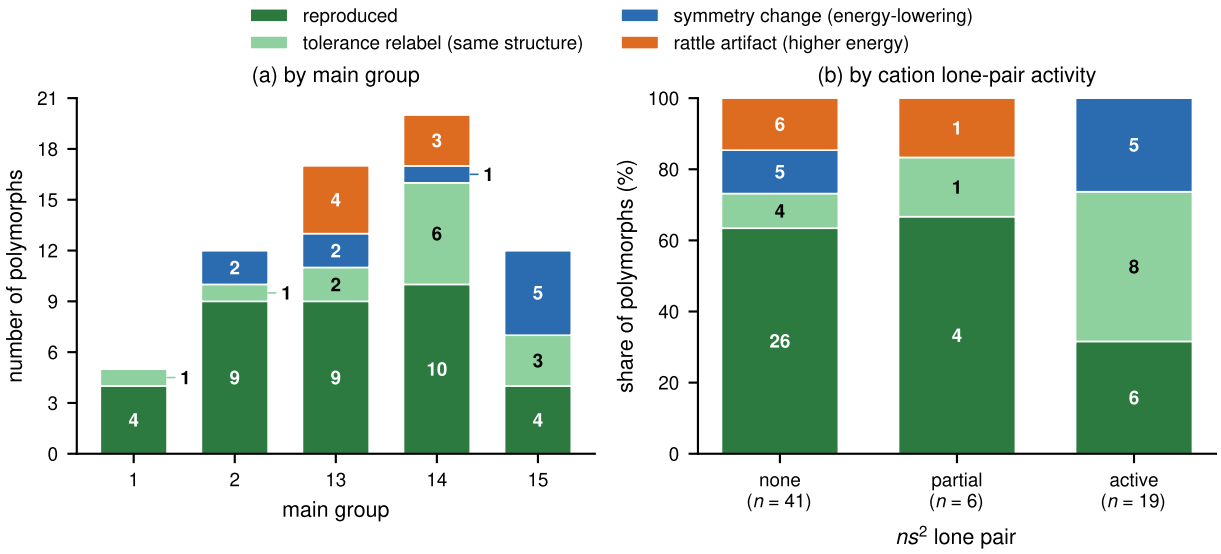


Figure S3: Outcome of the space-group reproduction audit under symmetry-breaking (rattle) relaxation, resolved (a) by main group and (b) by the cation ns^2 lone pair: active when every cation is in a lone-pair oxidation state (Tl^+, Sn^{2+}, Pb^{2+}, Bi^{3+}), partial for the mixed-valence hosts in which only the lower-valence cation carries a lone pair (Tl_4O_3, Pb_3O_4, Pb_2O_3, Bi_4O_7), and none otherwise. Panel (b) shows each outcome as a share of its class, with the class size n below each bar; segment labels give the number of polymorphs.

Figure S4 summarizes the audit energetics, plotting the rattle-minus-hs energy difference against the Materials Project energy above the convex hull for the 65 polymorphs with a Materials Project entry (*Pbca* SnO has none). The Materials Project hull is used because the r^{2}SCAN hull energies of the excluded idealized cells refer to their post-collapse structures. The six polymorphs excluded as unstable idealized cells for a primitive-cell instability relax more than 5 meV atom^{-1} below their idealized structures on rattling and lie more than 120 meV atom^{-1} above the hull (the other three members of the class, $P\bar{3}m1$ Bi_2O_3, $P2_1/m$ Pb_2O_3, and $Fm\bar{3}m$ PbO_2, were identified by the supercell checks below and in the Excluded polymorphs subsection above), whereas the two Al_2O_3 rattle artifacts discussed above appear as positive differences of 128 and 378 meV atom^{-1} at hull energies below 20 meV atom^{-1}.

One further check concerns the reference supercells themselves. The dense-matched pristine supercells (main-text Methods) are tiled from the relaxed cell of the vacancy-hosting stage and relax from that periodic tiling, so a distortion whose period exceeds the primitive cell cannot appear in them, and the primitive-cell rattle of Table S2 cannot reach it either. The $P\bar{3}m1$ polymorph of Bi_2O_3 is such a case, and it is excluded from the training set for it (Excluded polymorphs subsection above): its 90-atom reference converged in one ionic step at the symmetric configuration (maximum force 0.010 eV Å^{-1}), whereas in every defect cell of the

host at least 15 of the 18 interlayer oxygen atoms were displaced by 0.3–0.6 Å. A fixed-cell re-relaxation of the reference from a rattled start (Gaussian displacements with a standard deviation of 10 % of the shortest bond, the primitive-cell recipe without the lattice shear, at the same Γ-centered $3 \times 3 \times 3$ mesh) converged 1.16 eV (12.9 meV atom^{-1}) lower, with all 18 interlayer oxygen atoms displaced by 0.41 Å on average toward three of their six bismuth neighbors and no residual symmetry; since the re-relaxation is a single fixed-cell run from a random start, the 1.16 eV is a lower bound on the height of the symmetric configuration above the fixed-cell minimum.

The same check was then applied to the pristine reference supercell of every polymorph in the fit (Table S3). Of the 39 references, 32 returned to their starting structure, with no atom displaced by more than 0.1 Å after removal of the rigid shift and residual energy changes between −12 and +53 meV per supercell, within the force tolerance of the relaxations; none relaxed to a different structure of lower energy; and seven re-relaxations were trapped in different minima of higher energy, six of them, including Pb_3O_4, through oxygen pairs closer than 1.6 Å that formed during the relaxation (no starting structure had an O–O distance below 1.7 Å) and that a fixed-cell relaxation does not separate. The vacancy-induced displacement screen that exposed the $P\bar{3}m1$ Bi_2O_3 reference (root-mean-square host displacements of 0.29 and 0.23 Å at its two sites, against a median of 0.06 Å over all sites) gives at most 0.10 Å at any site of the hosts whose check was trapped, so none of them shows the signature of a symmetric-configuration reference.

The calculations used VASP 6.4: CPU builds of 6.4.1 for most calculations and a GPU build of 6.4.2 on a second cluster for part of them. The CPU and GPU builds of 6.4.2 on that cluster reproduce the 6.4.1 total energy of a 96-atom control supercell to within 4.2×10^{-4} meV atom^{-1}. Most of the rattled reference re-relaxations above, and all calculations of α-Bi_2O_3, ran on the GPU build. A third build, 6.4.2 for CPUs compiled in November 2023 and not included in the build control, ran the Γ-point configuration screening of a few group-14 vacancy sites; its energies enter only the ranking of configurations within a screening set, and every formation energy of the fit comes from a dense-matched pair computed with the 6.4.1 or the 6.4.2 GPU build. The electronic settings of the extended methods above apply to every calculation, with these exceptions: one dense-mesh defect relaxation (*Pbcn* In_2O_3, site 8) used the 1×10^{-6} eV self-consistency threshold of the band-gap calculations, and that relaxation, two other dense-mesh defect relaxations, and one band-gap calculation used blocked-Davidson minimization instead of the all-band algorithm.

Table S3: Rattled re-relaxation of the pristine reference supercell of each of the 39 polymorphs that contribute sites to the fit.

Group	Compound	Space group	N	ΔE (meV)	max $\lvert\Delta r\rvert$ (Å)	$n_{>0.1}$	Outcome[a]
1	Cs_2O	$R\bar{3}m$	96	3.8	0.099	0	same
1	K_2O	$Fm\bar{3}m$	108	−9.7	0.033	0	same
1	Li_2O	$Fm\bar{3}m$	240	−9.9	0.022	0	same
1	Na_2O	$Fm\bar{3}m$	192	23.8	0.025	0	same
1	Rb_2O	$Fm\bar{3}m$	108	5.0	0.054	0	same
2	BaO	$Fm\bar{3}m$	128	−3.4	0.017	0	same
2	BeO	$P4_2/mnm$	96	−11.6	0.015	0	same
2	BeO	$P6_3mc$	240	19.9	0.017	0	same
2	CaO	$Fm\bar{3}m$	96	0.7	0.005	0	same
2	MgO	$Fm\bar{3}m$	160	−4.2	0.009	0	same
2	MgO	$P6_3mc$	128	53.1	0.045	0	same
2	SrO	$Fm\bar{3}m$	96	−1.9	0.007	0	same
13	Al_2O_3	$C2/m$	160	2.5	0.007	0	same
13	Al_2O_3	$Pna2_1$	80	0.0	0.005	0	same
13	Al_2O_3	$R\bar{3}c$	180	0.4	0.004	0	same
13	Ga_2O_3	$C2/m$	160	12 325.9	4.860	105	higher, O–O pair
13	Ga_2O_3	$R\bar{3}c$	180	−0.5	0.008	0	same
13	In_2O_3	$Ia\bar{3}$	160	0.1	0.009	0	same
13	In_2O_3	$Pbca$	160	−0.8	0.012	0	same

continued on next page

Table S3 – continued

Group	Compound	Space group	N	ΔE (meV)	max $\lvert\Delta r\rvert$ (Å)	$n_{>0.1}$	Outcome[a]
13	In_2O_3	$Pbcn$	160	0.2	0.010	0	same
13	In_2O_3	$R\bar{3}c$	180	2.4	0.008	0	same
13	Tl_2O	$R\bar{3}m$	96	5.4	0.032	0	same
13	Tl_2O_3	$Ia\bar{3}$	160	1605.9	1.748	14	higher, O–O pair
13	Tl_4O_3	$P2_1/m$	84	−1.8	0.054	0	same
14	Pb_2O_3	$P2_1/c$	160	11 280.4	3.302	123	higher, O–O pair
14	Pb_3O_4	$P4_2/mbc$	224	91 212.2	3.981	204	higher, O–O pair[b]
14	PbO	$P4/nmm$	108	12.3	0.063	0	same
14	PbO	$Pbcm$	96	8.5	0.053	0	same
14	PbO_2	$P4_2/mnm$	216	−0.1	0.046	0	same
14	PbO_2	$Pbcn$	96	−0.8	0.017	0	same
14	SnO	$P4/nmm$	108	2596.6	5.889	42	higher
14	SnO	$Pbca$	96	4.1	0.039	0	same
14	SnO_2	$P4_2/mnm$	216	−9.2	0.009	0	same
15	Bi_2O_3	$P\bar{4}2_1c$	160	17 760.5	5.344	139	higher, O–O pair
15	Bi_2O_3	$P2_1/c$	160	−4.6	0.035	0	same
15	α-Bi_2O_3	$P2_1/c$	160	0.2	0.022	0	same
15	Bi_2O_3	$P31c$	160	8425.4	3.551	140	higher, O–O pair
15	Bi_2O_3	$Pccn$	80	−0.9	0.028	0	same
15	Bi_4O_7	$P\bar{1}$	176	−0.1	0.039	0	same

Each pristine supercell, relaxed at fixed cell on the production mesh, was perturbed by Gaussian displacements with a standard deviation of 10 % of its shortest interatomic distance and relaxed again at fixed cell on the same mesh. ΔE is the energy of the re-relaxed cell minus that of the reference, per supercell of N atoms; max $\lvert\Delta r\rvert$ is the largest atomic displacement from the reference after removal of the rigid shift, each atom paired with a reference atom of the same species by a minimum-cost (Hungarian) assignment, and $n_{>0.1}$ the number of atoms displaced by more than 0.1 Å. [a] same, the re-relaxed cell returns to the reference structure (no atom displaced by more than 0.1 Å), whatever the sign of the residual ΔE; lower, a different structure lower in energy by more than 10 meV per supercell; higher, a different structure higher in energy, a minimum in which the perturbed start was trapped (O–O pair: it contains an oxygen pair closer than 1.6 Å that the reference does not). [b] Stopped at the 48 h wall-time limit of its run after 22 ionic steps and not continued, its structure already holding O–O pairs; the last ionic step is reported.

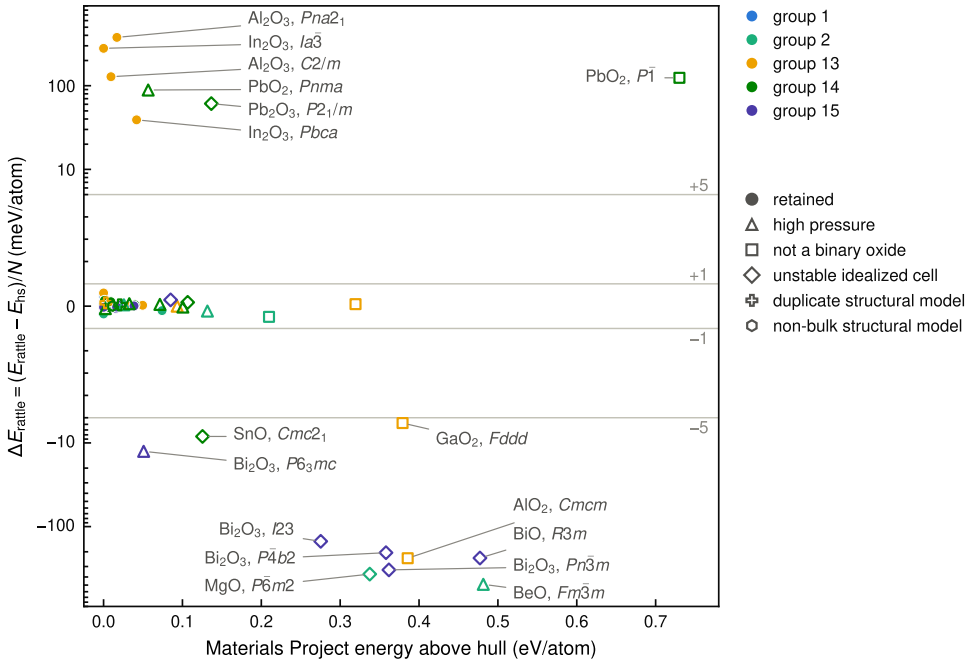


Figure S4: Rattle-minus-higher-symmetry (hs) energy difference per atom against the Materials Project energy above the convex hull for the 65 polymorphs with a Materials Project entry (symmetric logarithmic energy axis, linear within $\pm 5\,\mathrm{meV\,atom^{-1}}$; guides at ± 1 and $\pm 5\,\mathrm{meV\,atom^{-1}}$). Color encodes the main group and marker shape encodes the curation outcome; every point beyond $\pm 5\,\mathrm{meV\,atom^{-1}}$ is labeled.

Polymorph inventory and curation status

Table S4 enumerates all 66 material–polymorph combinations in the dataset (main-text Figure 1), labeled by the space group that the Materials Project assigns to each and by the structure type and mineral name recorded in the source ICSD entries, together with the number of ICSD collection codes that the Materials Project lists for each polymorph, whose records were screened for temperature and pressure (467 in total), and the training-set curation outcome. The 27 exclusions fall into the five classes of the Excluded polymorphs subsection of the extended methods above, which gives the per-phase evidence and the structural references.

The classification does not depend on the exact bounds of the near-ambient window. Any pressure bound from 0.2 MPa to 1 GPa gives the same classification: the only screened entries between these pressures belong to corundum Al_2O_3 (1.0 MPa) and Bi_2O_3-HP(II) (300 MPa), both of which have other near-ambient entries. Three polymorphs meet the pressure bound but not the temperature window: $P4_2/mnm$ BeO, retained on the structural grounds given in the Retained training set subsection above, and $P\bar{6}m2$ MgO and $P\bar{1}$ PbO_2, excluded as an unstable idealized cell and as a compound that is not a binary oxide. The temperature window therefore decides no curation outcome on its own.

Table S4: All 66 material–polymorph combinations in the dataset, ordered by the main group of the cation and labeled by the space group that the Materials Project assigns to each, with the ICSD structure type, the number of listed ICSD collection codes, and the training-set curation outcome.

Group	Compound	Space group	Structure type[a]	N_{ICSD}[b]	Curation[c]
1	Cs_2O	R$\bar{3}$m	$CdCl_2$ [anti]	1	retained
1	K_2O	Fm$\bar{3}$m	CaF_2	4	retained
1	Li_2O	Fm$\bar{3}$m	CaF_2	10	retained
1	Na_2O	Fm$\bar{3}$m	CaF_2	3	retained
1	Rb_2O	Fm$\bar{3}$m	CaF_2	3	retained
2	BaO	Fm$\bar{3}$m	NaCl	7	retained
2	BaO	P4/nmm	InBi	2	high pressure
2	BaO	P6$_3$/mmc	NiAs	1	high pressure
2	BeO	Fm$\bar{3}$m	NaCl	7	high pressure
2	BeO	P4$_2$/mnm	BeO (tP8)	2	retained
2	BeO	P6$_3$mc	ZnS [2H] (bromellite)	32	retained
2	CaO	Cmce	—	1	not a binary oxide
2	CaO	Fm$\bar{3}$m	NaCl (lime)	14	retained
2	MgO	Fm$\bar{3}$m	NaCl (periclase)	54	retained
2	MgO	P$\bar{6}$m2	—	1	unstable idealized cell
2	MgO	P6$_3$mc	CaO (hP4)	4	retained
2	SrO	Fm$\bar{3}$m	NaCl	12	retained
13	Al_2O_3	C2/m	Ga_2O_3	1	retained
13	Al_2O_3	Pbcn	Rh_2S_3	4	high pressure
13	Al_2O_3	Pna2$_1$	$AlFeO_3$	4	retained
13	Al_2O_3	R$\bar{3}$c	Al_2O_3 (corundum)	75	retained
13	AlO_2	Cmcm	FeOOH (boehmite)	15	not a binary oxide
13	AlO_2	Pnma	AlO(OH) (diaspore)	8	not a binary oxide
13	Ga_2O_3	C2/m	Ga_2O_3	5	retained
13	Ga_2O_3	R$\bar{3}$c	Al_2O_3	4	retained
13	GaO_2	Fddd	—	1	not a binary oxide
13	In_2O_3	Ia$\bar{3}$	Mn_2O_3	17	retained
13	In_2O_3	Pbca	Rh_2O_3	1	retained
13	In_2O_3	Pbcn	Rh_2S_3	4	retained
13	In_2O_3	R$\bar{3}$c	Al_2O_3	3	retained
13	Tl_2O	C2/m	—	1	duplicate structural model
13	Tl_2O	R$\bar{3}$m	—	1	retained

continued on next page

Table S4 – continued

Group	Compound	Space group	Structure type[a]	N_{ICSD}[b]	Curation[c]
13	Tl_2O_3	$Ia\bar{3}$	Mn_2O_3 (avicennite)	6	retained
13	Tl_4O_3	$P2_1/m$	—	1	retained
14	Pb_2O_3	$P2_1/c$	Pb_2O_3 (mP20)	1	retained
14	Pb_2O_3	$P2_1/m$	—	1	unstable idealized cell
14	Pb_3O_4	$P4_2/mbc$	Pb_3O_4 (minium)	7	retained
14	Pb_3O_4	Pbam	(minium)	1	high pressure
14	PbO	P4/nmm	PbO (litharge)	21	retained
14	PbO	Pbcm	PbO [orh] (massicot)	6	retained
14	PbO	$Pca2_1$	(massicot)	1	duplicate structural model
14	PbO_2	$Fm\bar{3}m$	CaF_2	1	unstable idealized cell
14	PbO_2	$P\bar{1}$	—	1	not a binary oxide
14	PbO_2	$P4_2/mnm$	TiO_2 (tP6) (plattnerite)	7	retained
14	PbO_2	Pbcn	α-PbO_2	6	retained
14	PbO_2	Pnma	$PbCl_2$	9	high pressure
14	SnO	$Cmc2_1$	—	1	unstable idealized cell
14	SnO	P4/nmm	PbO (romarchite)	5	retained
14	SnO	Pbca	—	1	retained
14	SnO	$Pmn2_1$	—	1	high pressure
14	SnO_2	$P4_2/mnm$	TiO_2 (tP6) (cassiterite)	42	retained
14	SnO_2	$Pa\bar{3}$	PdF_2 (cP12)	7	high pressure
14	SnO_2	Pbcn	—	1	non-bulk structural model
14	SnO_2	Pnnm	$CaCl_2$	3	high pressure
15	Bi_2O_3	I23	—	1	unstable idealized cell
15	Bi_2O_3	$P\bar{3}m1$	La_2O_3	1	unstable idealized cell
15	Bi_2O_3	$P\bar{4}2_1c$	β-Bi_2O_3	6	retained
15	Bi_2O_3	$P\bar{4}b2$	β-Bi_2O_3	2	unstable idealized cell
15	Bi_2O_3	$P2_1/c$	Bi_2O_3-HP(I) (relaxed)[d,e]	2	retained
15	α-Bi_2O_3	$P2_1/c$	α-Bi_2O_3[e]	10	retained
15	Bi_2O_3	P31c	Bi_2O_3-HP(II) (meta)[d]	4	retained
15	Bi_2O_3	$P6_3mc$	Bi_2O_3-HP(III) (meta)[d]	2	high pressure
15	Bi_2O_3	Pccn	Sb_2O_3	2	retained
15	Bi_2O_3	$Pn\bar{3}m$	Ag_2O_3	3	unstable idealized cell
15	Bi_4O_7	$P\bar{1}$	Bi_4O_7	1	retained
15	BiO	R3m	—	1	unstable idealized cell

[a] Prototype recorded in the retrieved ICSD CIFs of the polymorph (majority label when they differ), with the mineral name of the phase in parentheses where a strict majority of the retrieved CIFs record one, and an em dash where they record neither; the parenthesized Pearson symbols (tP8, hP4, mP20, tP6, cP12) and the bracketed polytype, modification, and antitype tags ([2H], [orh], [anti]) belong to the ICSD prototype label and are not mineral names; for the M_2O oxides the CaF_2 and $CdCl_2$ prototypes are antitypes, with cation and anion sites interchanged. [b] Number of ICSD collection codes that the Materials Project lists for the polymorph (467 in total; 76 are computed structures, which were not counted, and 17 could not be retrieved from the ICSD), whose records were screened for temperature and pressure. [c] Training-set curation outcome, with 39 polymorphs retained for crystal-feature-model fitting and 27 excluded in five classes: 10 high-pressure phases, 5 whose only structural determination is of a compound that is not a binary oxide, 9 whose computed cells are dynamically unstable, in the primitive cell or the vacancy supercell, or do not reproduce the recorded structure, 2 that are earlier structural models of phases retained under another space group, and 1 whose only structural record is a model of an interface layer, not of the bulk phase. [d] The Bi_2O_3-HP(I) (relaxed), HP(II) (meta), and HP(III) (meta) labels, qualifiers included, are the ICSD structure-type names of the multianvil high-pressure phases: HP(II) ($P31c$) is the quench-recovered phase, metastable at ambient conditions, and HP(I) ($P2_1/c$) is the intermediate to which it relaxes at ambient pressure; both are retained because their entries at 293–298 K and ambient pressure meet the near-ambient criterion (temperature within 273–313 K and pressure $\leq$0.2 MPa), whereas the HP(III) ($P6_3mc$) entries are in situ determinations at 2.8–5.4 GPa of a phase that reverts to HP(II) on decompression. [e] The two $P2_1/c$ Bi_2O_3 cells are Materials Project entries mp-1197128 (Bi_2O_3-HP(I)) and mp-23262 (α-Bi_2O_3, the mineral bismite, whose structure was determined by Malmros[S60]).

Main-text Figure 1 presents the same inventory as a composition–space-group coverage map, with the curation outcome encoded per cell; Figure S5 summarizes the structural diversity of the 66 polymorphs by

crystal system and lattice centering, and Figure S6 traces each main group's polymorphs to their curation outcomes.

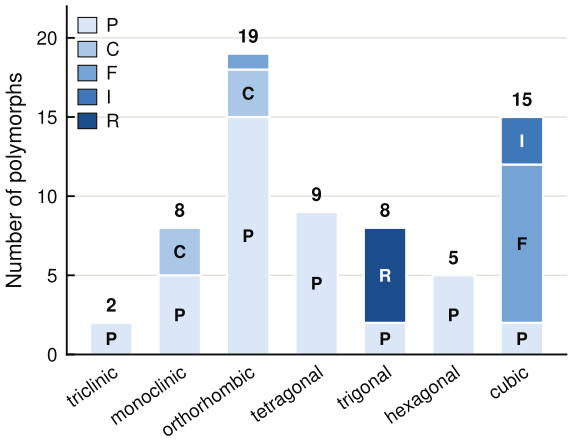


Figure S5: Crystal-system counts of the 66 polymorphs in the dataset, segmented by lattice-centering letter (P primitive, C base-centered, F face-centered, I body-centered, R rhombohedral), both read from the Materials Project space-group label (Table S4). Within the trigonal system, the centering letter, not the crystal-system label, fixes the Bravais lattice: cells with P centering belong to the hexagonal Bravais lattice, and those with R centering to the rhombohedral one.

Four of the retained compositions, with one retained polymorph each, are mixed-valence oxides, in which the same element occupies crystallographically distinct sites in two oxidation states: Tl_4O_3 (Tl^{+}/Tl^{3+}), Pb_3O_4 and Pb_2O_3 (Pb^{2+}/Pb^{4+}), and Bi_4O_7 (Bi^{3+}/Bi^{5+}). Figure S7 shows their relaxed structures with the cations colored by oxidation state. The states were assigned by bond-valence sums on the relaxed cells (computed as in PYMATGEN: R_0 from one set of atomic parameters per element,[S61] independent of oxidation state, $b = 0.31$ Å, and all neighbors within 4.0 Å), with the cations split into two groups in the ratio fixed by the stoichiometry; the two groups are separated by at least 2.5 valence units in every host, so the assignment is unambiguous. In all four structures, the high-valence cation is octahedrally coordinated, while the low-valence cation adopts the asymmetric coordination typical of the Tl^{+}, Pb^{2+}, and Bi^{3+} lone-pair ions. These are the hosts whose vacancy energetics motivate the oxidation-state-resolved local descriptors examined in the main-text outlier analysis.

Figure S8 documents the structural evidence behind one of the exclusions, the $Cmc2_1$ polymorph of red SnO: the lowest-energy configuration of its oxygen vacancy is not a local rebonding but a global collapse of the host toward β-tin-like Sn–Sn bonding, consistent with the superseded status of the underlying structure determination (Excluded polymorphs subsection above and main-text outlier analysis). The *Pbca* cell of red SnO from the single-crystal redetermination, computed with the identical workflow and retained in the training set, shows none of this pathology (main-text outlier analysis): its rattled copy relaxes to a configuration within 1 meV atom^{-1} of it (Table S2), its single symmetry-inequivalent vacancy relaxes without host reconstruction (the full SHAKENBREAK distortion set spans 5 meV), and its formation energy of 4.08 eV is underpredicted by the primary model by 0.4 eV, on the same side as litharge SnO.

A parallel collapse occurs in the fluorite-type ($Fm\bar{3}m$) polymorph of PbO_2, whose only ICSD record (77648) is that of a phase synthesized at high pressure, and led to its exclusion (Excluded polymorphs subsection above). The ideal fluorite cell is a symmetry-protected stationary point, so its pristine supercell retains the ideal geometry on relaxation; the vacancy, however, triggers a global reconstruction in which every Pb drops from eightfold to sixfold oxygen coordination, placing the defect cell 5.14 eV below the pristine reference (a non-spin-polarized cell, like the defect cell) and yielding the formation energy of −11.07 eV (Figure S9). Re-relaxed at fixed cell from rattled starts (0.05 and 0.10 times the shortest interatomic distance), the pristine 96-atom supercell itself falls 69 and 96 meV atom^{-1} below the fluorite cell, toward structures near $I4_1/a$ and $Pca2_1$, without forming O–O pairs closer than 2.66 Å: the fluorite cell is not a

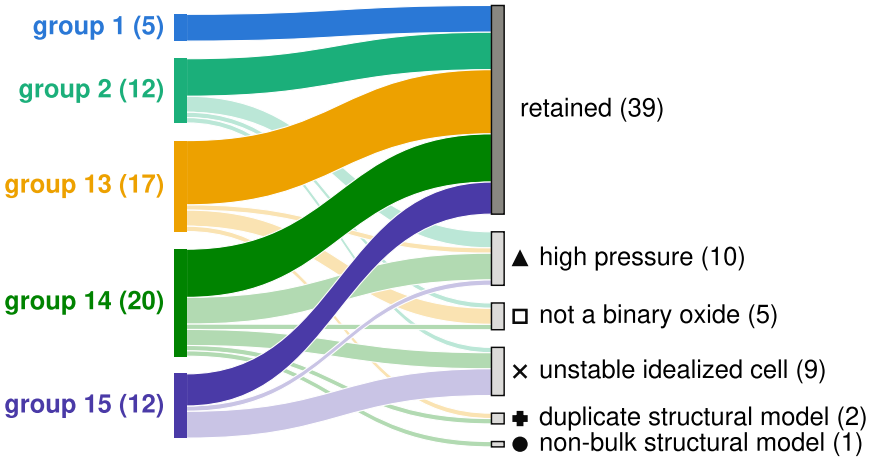


Figure S6: Curation of the 66 polymorphs in the dataset, from main group (left) to training-set outcome (right). Band widths are numbers of polymorphs; the 39 retained polymorphs train the crystal feature model, and the 27 excluded ones fall into the five classes of the main-text Polymorph Selection subsection (Table S4). As in main-text Figure 1, bands into the retained set take the solid main-group colors, bands into the exclusion classes their lighter tints, and each exclusion class carries that figure's marker.

minimum.

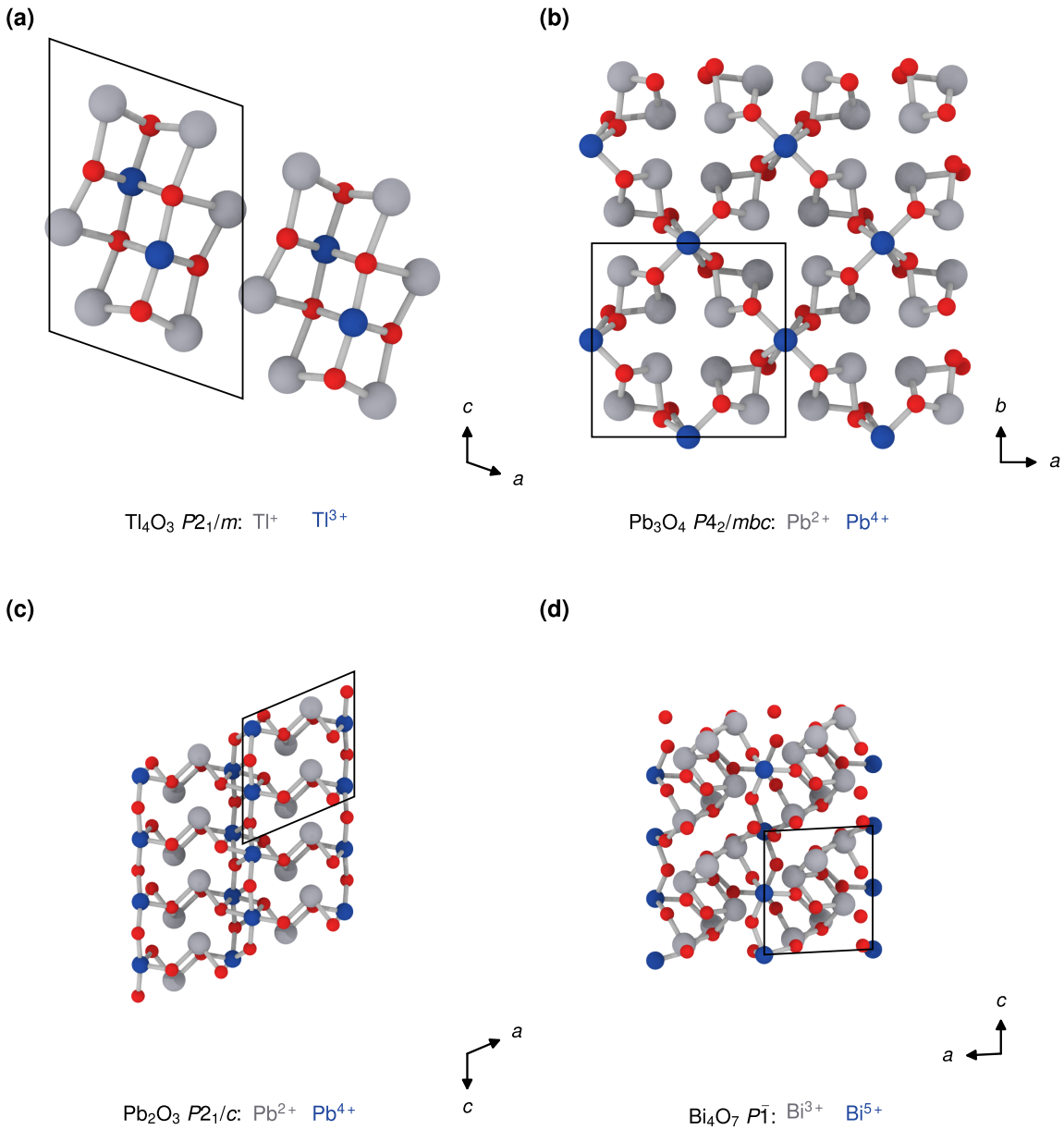


Figure S7: Relaxed structures of the four mixed-valence compositions in the training set (one polymorph of Pb_2O_3 shown), with cations colored by oxidation state (low-valence light gray and larger, high-valence dark blue and smaller) and oxygen red; the black outline marks one unit cell. (a) Tl_4O_3 ($P2_1/m$): ribbons of $Tl^{3+}O_6$ octahedra sheathed by Tl^{+}; the ribbons are linked only by longer Tl^{+}–O contacts (> 3 Å) that are not drawn. (b) Pb_3O_4 (minium, $P4_2/mbc$), viewed along c, the fourfold axis: chains of edge-sharing $Pb^{4+}O_6$ octahedra linked by pyramidally coordinated Pb^{2+}. (c) Pb_2O_3 ($P2_1/c$). (d) Bi_4O_7 ($P\bar{1}$). Metal–oxygen bonds are drawn below host-specific cutoffs of 2.35–2.85 Å chosen at the first-shell gap of each host's bond-length distribution.

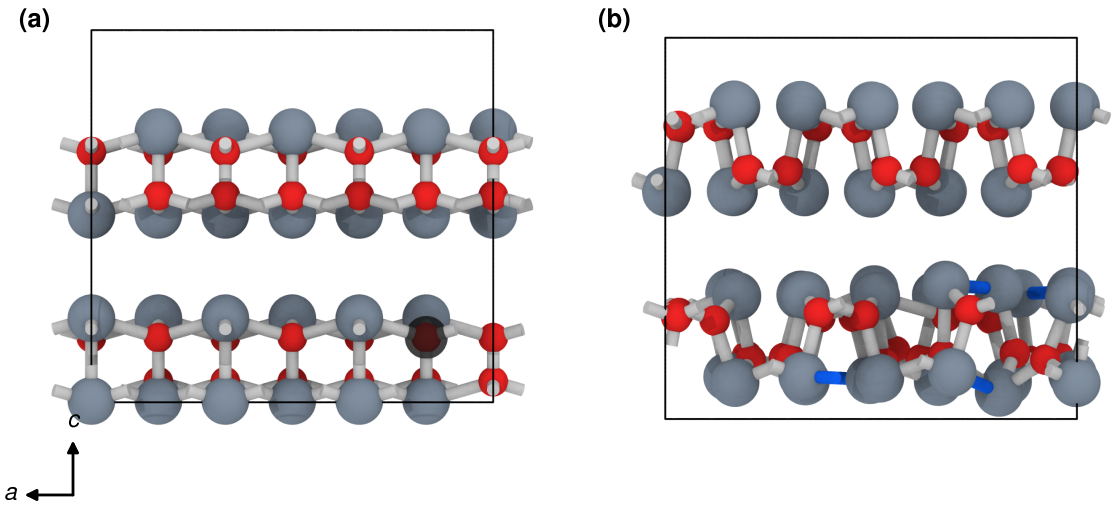


Figure S8: Vacancy-triggered collapse of the excluded $Cmc2_1$ polymorph of red SnO (Sn gray, O red; Sn–O bonds drawn below 2.6 Å; the black outline is the supercell boundary). (a) The relaxed pristine 96-atom supercell: all 48 Sn are four-coordinate, and the black sphere marks the oxygen removed to form the vacancy. (b) The ShakeNBreak ground state of that vacancy (site s6, +50 % bond distortion): 44 of 48 Sn become three-coordinate and short Sn–Sn contacts form (blue, drawn below 3.30 Å; shortest 3.09 Å, compared with 3.05 and 3.20 Å in β-tin relaxed at the same level of theory); panel (b) is drawn with its origin shifted by $0.10\,c$ so that the collapsed layer is not split by the cell boundary.

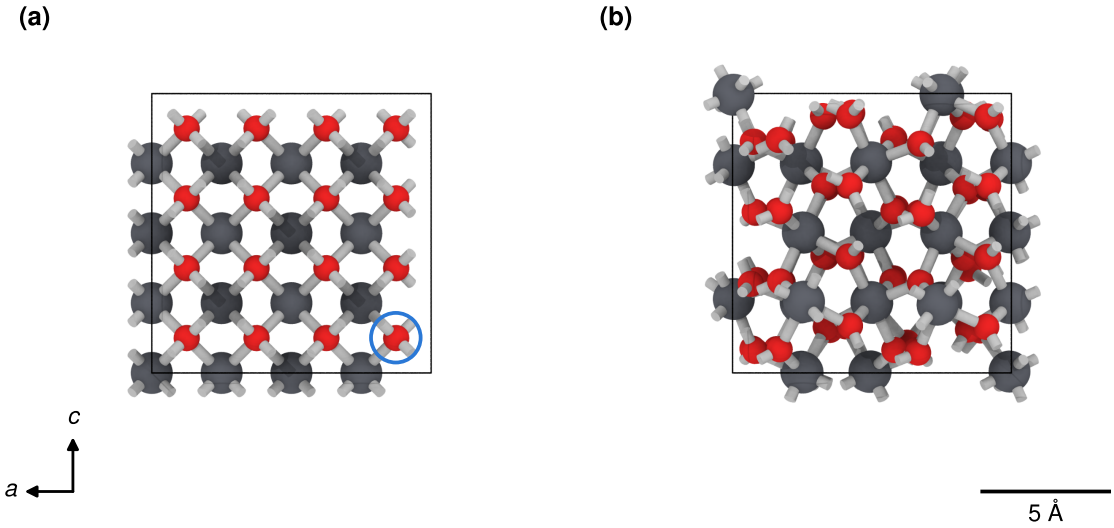


Figure S9: Vacancy-triggered collapse of the fluorite-type PbO_2 cell, excluded as an unstable idealized cell (Pb gray, O red; Pb–O bonds drawn below 2.6 Å; the black outline is the supercell boundary; both panels are drawn at the same scale). (a) The relaxed pristine 96-atom supercell, which retains the ideal fluorite geometry (all 32 Pb eight-coordinate); the blue ring marks the position of the oxygen removed to form the vacancy, projected along b (the other oxygen atoms of its column project onto the same position). (b) The ShakeNBreak ground state of that vacancy (+30 % bond distortion): the entire host reconstructs, every Pb becoming six-coordinate, and the defect cell falls 5.14 eV below the pristine supercell, yielding the vacancy formation energy $E_v = -11.07$ eV.

Single-descriptor correlations and model robustness

Figure S10 shows how each crystal-feature-model descriptor correlates with E_v in isolation over the 72 clean sites (those with a well-defined E_v, main-text Methods). The crystal reduction potential is the strongest single predictor (Pearson $R = -0.98$): the more readily a host's cation is reduced, the lower the vacancy formation energy. The band gap is the second strongest descriptor and correlates positively ($R = +0.88$), reflecting the additional cost of promoting electrons into the conduction band when cation reduction alone cannot accommodate them. The bond-dissociation sum ΣE_b correlates positively and moderately ($R = +0.76$); it correlates more strongly than the mean bond energy $\langle E_b \rangle$ ($R = +0.51$) because $\Sigma E_b = \mathrm{CN_O} \cdot E_b$ carries the site-specific coordination number, whereas $\langle E_b \rangle$ is a per-composition constant. The energy above the hull correlates only weakly as a lone predictor, and with the opposite sign to the one the Born–Haber cycle predicts ($R = +0.30$). The r^{2}SCAN-consistent and experimental-thermochemistry routes give nearly the same correlation for V_r ($R = -0.98$ and -0.97) and similar ones for ΣE_b ($+0.76$ and $+0.80$).

Both routes assign each composition its adjacent couple (main-text Methods). In the experimental route, the adjacent couple of Tl_4O_3, $Tl_4O_3 \rightarrow Tl_2O$, is 0.05 V less favorable than the reduction to the metal, a comparison that rests on the estimated enthalpy of formation of Tl_4O_3; in the r^{2}SCAN route, the adjacent couple of Tl_4O_3 is also its most favorable one. The grouped cross-validation of the main text pools all held-out sites; Table S5 lists the leave-one-composition-out error of each held-out composition.

Table S5: Per-composition errors of the primary two-descriptor model under leave-one-composition-out cross-validation: for each composition, the mean absolute error of its sites when the model is trained without it, for the room-temperature-stable model (RT) and the all-site model, with the number of sites (n) in each. The main-text values pool all held-out sites (RT 0.31 eV, all-site 0.29 eV); the means over compositions are 0.38 and 0.40 eV.

Composition	n (RT)	MAE, RT (eV)	n (all)	MAE, all (eV)
Cs_2O	1	0.39	1	0.48
K_2O	1	0.41	1	0.33
Li_2O	1	0.23	1	0.25
Na_2O	1	0.25	1	0.17
Rb_2O	1	0.40	1	0.31
BaO	1	0.40	1	0.51
BeO	2	0.35	2	0.14
CaO	1	0.24	1	0.33
MgO	1	0.01	2	0.26
SrO	1	0.10	1	0.22
Al_2O_3	4	0.16	10	0.18
Ga_2O_3	4	0.38	4	0.36
In_2O_3	2	0.23	7	0.18
Tl_2O	1	0.48	1	0.55
Tl_2O_3	1	1.34	1	1.40
Tl_4O_3	3	0.07	3	0.12
Pb_2O_3	3	0.38	3	0.37
Pb_3O_4	2	0.60	2	0.59
PbO	2	0.21	2	0.20
PbO_2	2	0.94	2	0.98
SnO	–	–	2	0.62
SnO_2	1	0.72	1	0.64
Bi_2O_3	13	0.16	15	0.16
Bi_4O_7	7	0.30	7	0.30

The instability of the hull coefficient in the four-descriptor fit (main text) reflects redundancy rather than missing physics. A signed counterpart of E_{hull}, the per-oxygen energy of reduction to the adjacent oxygen-poorer phases on our r^{2}SCAN convex hull, correlates with E_v as strongly as V_r itself ($R = +0.97$

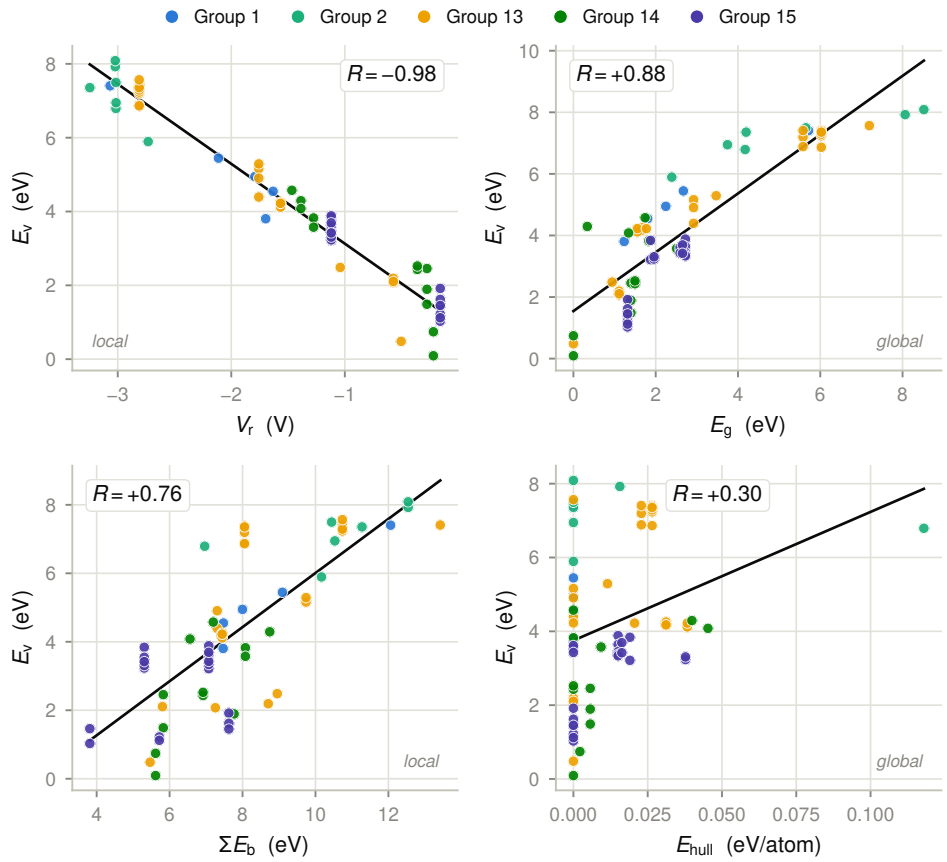


Figure S10: Single-descriptor Pearson correlations R of the four crystal-feature-model descriptors with the vacancy formation energy E_v over the 72 clean sites. The crystal reduction potential V_r and the bond-dissociation sum ΣE_b, both from the r^{2}SCAN-consistent route, are local descriptors (properties of the vacant oxygen's coordination environment, of which V_r takes one value per composition in a single-cation oxide); the band gap E_g and the energy above the hull E_{hull} are global (properties of the host crystal). Each point is one site, colored by main group; the line in each panel is the least-squares fit.

over the room-temperature-stable sites, against -0.97 for V_r), but it is collinear with V_r ($R = -0.999$ over all 72 clean sites, with a slope of -2.0 set by the two electrons transferred per released oxygen), and substituting it for E_hull does not lower the room-temperature-stable test MAE (0.35 against 0.33 eV on the same splits). E_hull therefore contributes only the polymorph-level metastability that V_r does not already encode; within the room-temperature window, that remainder spans 0–23 meV atom^{-1}, too narrow a range to constrain a coefficient.

Figure S11 details the sensitivity of the room-temperature-stable test mean absolute error (MAE) to the methodological choices that most affect the target energies or the band-gap descriptor: the pristine reference, the band-gap definition, and the production k-sampling. Every rung is fitted in the four-descriptor form on the same 1000 paired splits of the 56 room-temperature-stable sites common to all rungs, and each step is reported as the paired mean change of the test MAE with the share of splits in which it decreases. A Γ-only perfect supercell gives 0.42(7) eV: at that stage, the defect ground states of 22 sites (in 13 polymorphs) had been re-relaxed on a k-point mesh, so their defect and pristine energies were sampled differently. The relaxed bulk-cell energy multiplied by the number of bulk cells in the supercell gives 0.43(7) eV (+0.006 eV, lower in 42 % of the splits); a perfect supercell matched to each defect calculation in both cell and k-point mesh gives 0.43(8) eV, unchanged within the split-to-split spread (+0.007 eV, lower in 44 %); replacing the Materials Project legacy generalized-gradient band gap with the r^{2}SCAN fundamental gap gives 0.42(7) eV (-0.017 eV, lower in 86 %); and re-relaxing every defect ground state and its pristine reference at the dense production mesh of Table S1 gives the final convention, 0.33(5) eV (-0.091 eV, lower in 98 %; 0.31(4) eV for the two-descriptor model). The last step is a Brillouin-zone-sampling effect: the formation energies move by a median of 0.01 eV, but by up to 1.7 eV in the nine hosts with a site that moves by more than 0.5 eV (Bi_2O_3 $P\bar{4}2_1c$, In_2O_3 $Ia\bar{3}$, In_2O_3 $Pbca$, In_2O_3 $Pbcn$, In_2O_3 $R\bar{3}c$, PbO_2 $P4_2/mnm$, PbO_2 $Pbcn$, Tl_2O_3 $Ia\bar{3}$, and Tl_2O $R\bar{3}m$), whose defect states are dispersive.

The last step is not a change of spin state: three Γ-only pristine references, of $Ia\bar{3}$ Tl_2O_3, $Pbcn$ PbO_2, and the excluded fluorite-type PbO_2, had relaxed to 2 μ_B against closed-shell defect cells, and every rung uses instead their non-spin-polarized relaxations, which lie 27.6 meV above and 23.4 meV below the spin-polarized cells of the two retained hosts. The dense-matched pair is adopted on the physical grounds that it converges the Brillouin-zone integration of both supercells rather than relying on the cancellation of two sparse-sampling errors, not because it minimizes the cross-validated error.

Varying the stability window over which the model is trained shows that the reported accuracy is not an artifact of the window choice. As the E_hull cutoff is relaxed from 25 to 50 to 100 meV atom^{-1} and finally to no cutoff at all, the within-window test MAE of the primary model takes the values 0.31, 0.28, 0.28, and 0.29 eV (the 50 meV atom^{-1} window admits 15 additional near-window sites), while the MAE of the model refit on each whole window, evaluated on the fixed 56-site room-temperature core (a training-set, not held-out, error), remains at 0.28–0.29 eV. The window therefore neither inflates the reported accuracy nor is required for it; the 25 meV atom^{-1} restriction is retained for comparability with the perovskite parent. The full sweep is tabulated in the accompanying machine-readable data files.

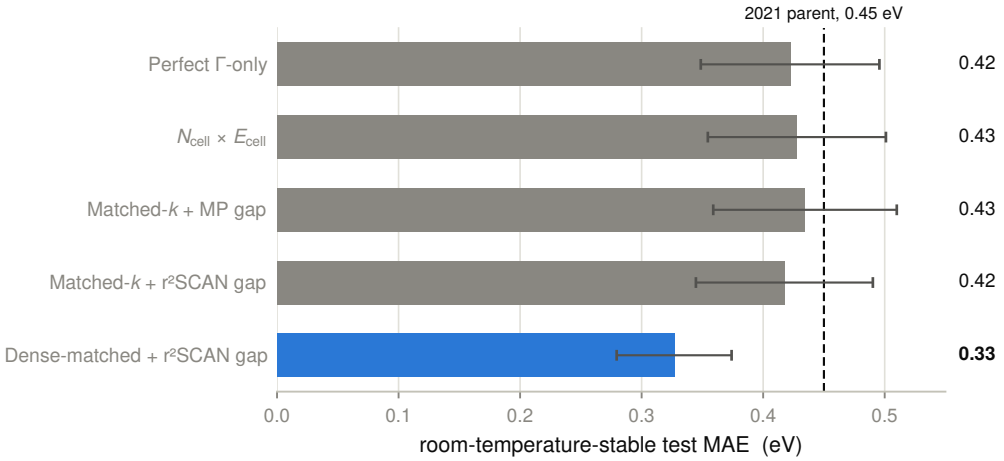


Figure S11: Sensitivity of the room-temperature-stable test MAE to the pristine-reference, band-gap, and production-sampling choices, evaluated on the four-descriptor form. Error bars are the standard deviation over 1000 random train–test splits, the same splits for every rung. Top to bottom, the pristine reference E_{bulk} of main-text Eq. 1 is moved from a Γ-only perfect supercell (whose sampling differs from that of the 22 defect ground states then re-relaxed on a k-point mesh) to the relaxed bulk-cell energy E_{cell} times the number N_{cell} of such cells in the supercell (the multiplicity of Table S1) and then a perfect supercell matched to each defect calculation in cell and k-point mesh (matched k); the Materials Project (MP) band gap is replaced with the r^{2}SCAN fundamental gap; and each defect ground state and its pristine supercell are re-relaxed at the dense production mesh rather than at their earlier sampling (Γ only, or a sparse mesh for those 22 sites; bottom, 0.33(5) eV, the adopted convention). The dashed line marks the 0.45 eV of the 2021 perovskite parent model.

Comparison of r^{2}SCAN and Materials Project hull energies

Figure S12 shows the r^{2}SCAN energy above the convex hull for all 39 retained polymorphs, grouped by composition in the ordering of main-text Figure 2, with two Materials Project values paired to each point (database version 2026.09.28): the default hull, which combines r^{2}SCAN with generalized-gradient (GGA and GGA + U) energies and whose composition offsets carry the database's GGA oxide correction, and the uncorrected r^{2}SCAN-only hull. Among the room-temperature-stable hosts, the three hulls differ by at most 6 meV atom^{-1}, consistent with the agreement quoted in the main-text Methods. The two database hulls differ for three polymorphs only: litharge SnO lies on the default hull but 40 meV atom^{-1} above the r^{2}SCAN-only one (40 meV atom^{-1} on ours), so its position on the default hull is a property of the mixing scheme rather than of the r^{2}SCAN energies, and Rb_2O and $P2_1/c$ Pb_2O_3 differ by at most 4 meV atom^{-1}.

Two polymorphs change window membership with the hull source: κ-Al_2O_3 ($Pna2_1$) lies 26 meV atom^{-1} above our hull, outside the window, and 21 meV atom^{-1} above both database hulls, inside it, and litharge SnO lies inside the window on the default database hull only; θ-Al_2O_3 ($C2/m$) lies inside it under all three (23 meV atom^{-1} on ours). Drawing the window on the default database hull adds these two hosts (63 sites) and gives the primary model a test MAE of 0.30 eV, and drawing it on the r^{2}SCAN-only hull gives 0.29 eV. The three hulls agree on the Bi_2O_3 ground state: α-Bi_2O_3 (bismite, $P2_1/c$) is the vertex in each, and the $P2_1/c$ Bi_2O_3-HP(I) polymorph of Table S4 lies 15 meV atom^{-1} above it on ours and 16 meV atom^{-1} on both database hulls. The 27 excluded polymorphs are omitted because several of their idealized cells relax to different structures, so their computed hull energies would not describe the nominal phases.

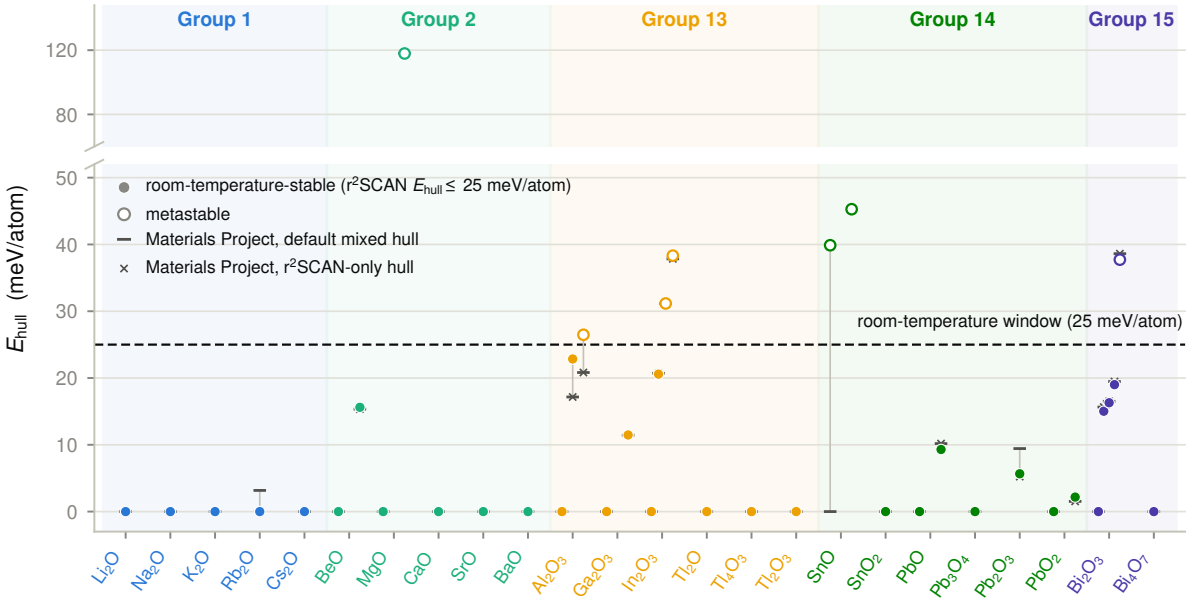


Figure S12: Energy above the r^{2}SCAN convex hull for the 39 retained polymorphs, one marker per polymorph, grouped by composition in the ordering of main-text Figure 2 and colored by main group. Filled symbols mark room-temperature-stable polymorphs ($E_{\mathrm{hull}} \leq 25$ meV atom^{-1}, dashed line); open symbols mark metastable polymorphs; the vertical axis is broken above 52 meV atom^{-1}. Gray dashes and crosses give the Materials Project default and r^{2}SCAN-only hull values (database version 2026.09.28) for the same polymorph, joined to the r^{2}SCAN value by a thin vertical line where they differ; *Pbca* SnO has no database entry.

Search for a site-resolved fifth descriptor

This section gives in full the search that the main text summarizes under Within-Host Degeneracy and the Search for a Site-Resolved Descriptor.

The two-descriptor model is exactly flat *within* each polymorph, because both of its descriptors are host-level quantities (in the four-descriptor form, site-to-site discrimination appears only through ΣE_b, whose null coefficient suppresses it). The observed within-host spreads are substantial: the seven sites of Bi_4O_7 span 0.89 eV, the three sites of monoclinic Ga_2O_3 span 0.77 eV, and the three sites of the mixed-valence Pb_2O_3 span 0.97 eV, each set at a single predicted value, so this descriptor degeneracy sets a floor on the within-host resolution of any model built from these quantities. We therefore searched systematically for a fifth, site-resolved descriptor (Figure S13). The outcome, developed below, is that one candidate, the shortest metal–oxygen bond length of the vacant site, d_{min}, correlates positively with E_v within each of the nine hosts with at least three sites, yet barely improves the model, because the magnitude of the within-host trend differs from host to host.

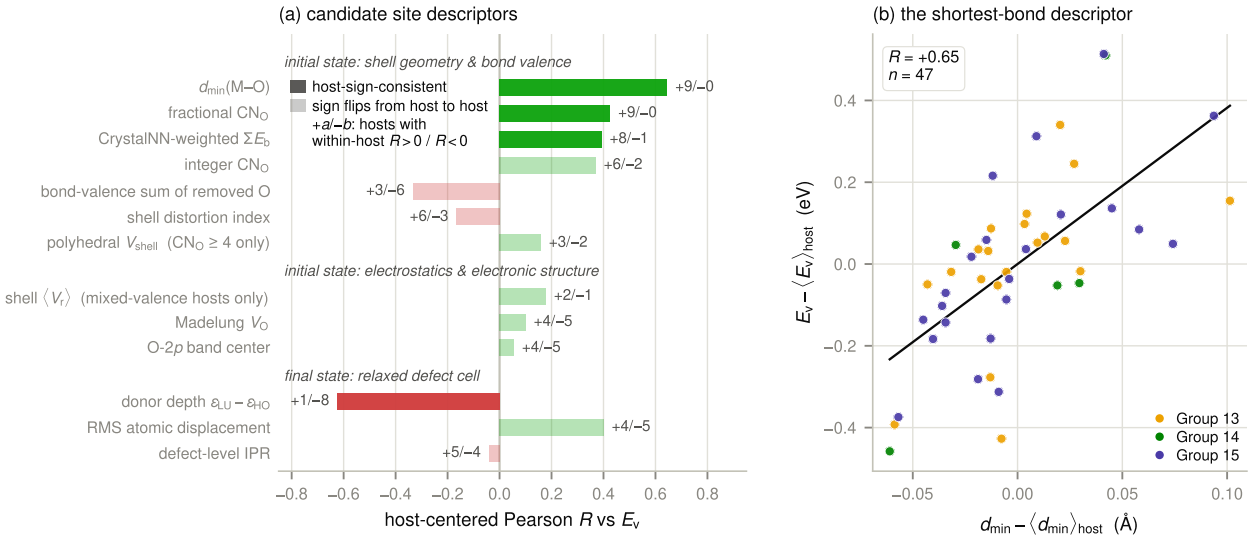


Figure S13: Candidate fifth descriptors tested against the variation of E_v among the sites of a single host. (a) Pearson correlation of each candidate site descriptor with E_v after each host's mean has been subtracted from both (host-centered values), pooled over the 47 sites of the 14 polymorphs with at least two sites (19 for the polyhedral shell volume V_{shell}, which requires at least four cation neighbors, $CN_O \geq 4$, that are not all coplanar; and 15 for the shell $\langle V_r \rangle$, which varies within a host only in the four polymorphs of the four mixed-valence compositions). The $+a/-b$ annotations count, among the hosts with at least three sites at which the descriptor is defined and varies, those whose own correlation is positive (a) or negative (b). Saturated bars mark sign-consistent candidates, whose per-host sign agrees with the pooled sign in at least 80 % of the counted hosts, among at least three counted hosts; pale bars fall short of that threshold and therefore cannot carry a single global coefficient. In the row labels, CN_O is the oxygen coordination number, V_O the point-ion Madelung potential at the vacant site, $\langle V_r \rangle$ the cation-shell average of the oxidation-state-resolved reduction potential, $\varepsilon_{LU} - \varepsilon_{HO}$ the Kohn–Sham donor depth, and IPR the inverse participation ratio of the defect level. (b) The sign-consistent candidate with the largest pooled correlation, the shortest metal–oxygen bond length d_{min} of the vacant site, against E_v, both host-centered.

Each candidate was judged on host-centered values: the mean over a polymorph's sites is subtracted from both the descriptor and E_v, leaving only site-to-site differences (47 points from the 14 polymorphs with at least two sites; evaluation conventions in the final section). A single global coefficient can describe a within-host trend only if that trend has the same sign in every host, so each candidate's correlation was also counted host by host, over the hosts with at least three sites (the $+a/-b$ annotations of Figure S13a). The initial-state candidates, properties of the pristine crystal available before any defect calculation, were the oxidation-state-resolved reduction potentials of the vacancy's cation shell, integer and fractional (CrystalNN-weighted)

oxygen coordination numbers CN_O, the CrystalNN-weighted bond-dissociation sum, the bond-valence sum of the removed oxygen, the point-ion Madelung potential at the vacant site, the volume and distortion index (bond-length regularity) of the polyhedron formed by the cation shell (the binary-oxide analogue of the perovskite octahedral volume), and the site-projected O-2p band center. The final-state candidates, extracted from the relaxed defect supercells, were the Kohn–Sham donor depth $\varepsilon_{LU} - \varepsilon_{HO}$ (the separation between the lowest unoccupied and the highest occupied eigenvalues of the defect supercell), the localization of the defect level (its inverse participation ratio, IPR), and the root-mean-square (RMS) displacement of the atoms of the relaxed defect supercell from the nearest pristine site of the same species.

Most candidates fail not because their correlations are weak but because their signs are inconsistent across hosts: the Madelung potential, the bond-valence sum, the RMS displacement, the O-2p band center, the shell distortion index, the polyhedral volume, and the defect-level IPR all reverse sign from host to host (the O-2p band center, for instance, correlates at $R = +0.84$ with E_v over the six sites of Bi_2O_3-HP(I) but at $R = -0.67$ over the seven sites of Bi_4O_7), so a global coefficient fit to one subset of hosts mispredicts the rest. The polyhedral volume is, moreover, undefined at the sites whose two- or three-coordinate oxygen bounds no volume, a structural limitation absent in the guaranteed octahedra of the perovskites (polyhedral-volume section below, Table S6).

Among the sign-consistent candidates, d_{min} is the strongest: its host-centered $R = +0.65$ stays between $+0.60$ and $+0.67$ when any one host is left out, and its correlation is positive in each of the nine hosts with at least three sites (Figure S13a); the fractional coordination number and the CrystalNN-weighted bond-dissociation sum, which also pass, add nothing once d_{min} is controlled for (partial $|R| \leq 0.02$). Within a host, the overbonded oxygen sites, those with the shortest bonds, are systematically the cheaper vacancies. This local strain-relief signature runs opposite to the naive bond-breaking expectation encoded in ΣE_b, and the bond-valence sum of the removed oxygen carries the same information with the expected opposite pooled sign (centered $R = -0.33$), although its per-host sign is mixed (negative in six of the nine hosts with at least three sites; Figure S13a). The bond length itself, not a proxy, carries the signal: within hosts, with the integer CN_O, $R(E_v, d_{min} \,|\, CN_O) = +0.57$ whereas $R(E_v, CN_O \,|\, d_{min}) = +0.05$, so the coordination number is a proxy for the bond length, not the reverse. Referenced to a natural metal–oxygen bond-length scale, d_{min} also correlates with E_v across hosts with the same sign as within them, and the within-host slope is of the order of the natural force scale of a metal–oxygen bond (shortest-bond section below), so the trend reflects local bond compression rather than cation size. This ordering does not, however, translate into accuracy: adding d_{min}, centered on its mean over all oxygen sites of the host's bulk cell, to the primary model changes its held-out test MAE on the room-temperature-stable sites by only -0.024 eV, with the 2.5th to 97.5th percentiles of the paired difference over the 1000 train–test splits spanning -0.055 eV to $+0.026$ eV.

The search stops there for three reasons. First, a site-resolved reduction potential makes the model worse. Replacing the composition-level V_r by an oxidation-state-resolved average over the vacancy's first cation shell (which differs from the composition value only in the mixed-valence hosts) *degrades* the four-descriptor model (test MAE 0.33 eV $\rightarrow$ 0.48 eV on the room-temperature-stable sites and 0.30 eV $\rightarrow$ 0.44 eV on all 72 sites). The local and global trends need not share a sign: across compositions, a more reducible cation lowers E_v, but in Pb_2O_3, where Pb^{4+} is by far the more reducible of the two cations, vacancies coordinated by more Pb^{4+} are *costlier* to form (Figure S7 shows the four mixed-valence structures with their cation oxidation states resolved). The stronger binding of oxygen to its higher-charged neighbors (shorter Pb^{4+}–O bonds, a deeper electrostatic well at the site) outweighs the greater reducibility of the shell. The host-level V_r is thus the appropriate formulation: the adjacent reduction couple available to the two released electrons is a property of the crystal as a whole, not an average over the vacancy's first coordination shell.

Second, outside the bond-geometry family (the coordination-number and bond-sum rows are proxies for d_{min}, as the partial correlations above show), the Kohn–Sham donor depth is the only sign-consistent candidate, negative in eight of the nine hosts, but it overlaps with d_{min}: its host-centered correlation with E_v weakens from $R = -0.63$ to a partial $R = -0.45$ once d_{min} is controlled for, and in leave-one-host-out tests (one polymorph held out at a time), the two together recover 44 % of the host-centered variance (50 % in the room-temperature window), against 39 % (39 %) for d_{min} alone. Third, no single descriptor from a published library[S62] passes the false-discovery-rate, sign-consistency, and leave-one-host-out tests; the within-host signal that remains beyond d_{min} is spread over many fingerprints (shortest-bond section below).

The final-state electronic structure limits what any site-resolved descriptor can recover, which is why the

search saturates, and at the same time confirms the Born–Haber picture the model is built on. The neutral vacancy forms no polaron at any site (the largest site-projected moment is $0.03\,\mu_{\mathrm{B}}$), and its defect level is a single doubly occupied band at 70 of the 72 sites; the two electrons are spin-paired except in the two defect cells that keep a net moment (extended methods above), the largest that of BaO ($1.06\,\mu_{\mathrm{B}}$, of which $0.25\,\mu_{\mathrm{B}}$ lies within the atomic projection spheres). In the gapped hosts, the level lies a median of $0.6\,E_{\mathrm{g}}$ below the conduction edge, as estimated from the eigenvalues of the defect supercell (an upper bound on the depth, since that estimate exceeds the bulk gap at a few sites), deep at all but four sites (one each in BaO $Fm\bar{3}m$, Bi_2O_3 $P\bar{4}2_1c$, In_2O_3 $Pbcn$, and Tl_2O $R\bar{3}m$, where it is within $0.1\,E_{\mathrm{g}}$ of the conduction edge); in the three metallic hosts it merges with the conduction states. At 22 of the 72 sites, the metal component of the level, projected from the full-precision `vasprun.xml` output, is predominantly s-like (more than half of the metal weight; 31 by plurality), the cation-reduction step of the cycle made directly visible. The site-to-site ordering of E_{v} thus tracks the depth of this defect level, which in turn tracks the local bond geometry that d_{min} encodes, but the *magnitude* of that dependence is host-specific. The remaining within-host variation (an RMS of 0.18 eV in the room-temperature window and 0.16 eV over all hosts after a pooled d_{min} term fitted to these same sites, a lower bound for any single global coefficient) is therefore irreducible for any single global linear term and points instead to hierarchical, per-host-calibrated extensions.

Polyhedral volume and regularity of the vacancy coordination shell

In perovskites, where every oxygen is a shared corner of two BO_6 octahedra of fixed topology, the octahedral volume is a natural and always defined geometric descriptor of the vacancy site. The analogous construction in a binary oxide is the polyhedron spanned by the vacancy's first-shell cations. For each of the 72 clean sites, we computed four measures on the pristine relaxed supercell with the same CrystalNN convention [S56] that defines CN_O and ΣE_b in the model: the convex-hull volume of the cation shell V_{shell}; its dimensionless normalized form $V_{shell}/\langle d\rangle^3$; the Baur bond-length distortion index $DI = \langle|d_i - \langle d\rangle|\rangle/\langle d\rangle$; and the off-centering of the oxygen from the shell centroid, normalized by $\langle d\rangle$.

In the binary oxides, the perovskite premise of a guaranteed polyhedron does not hold: 27 of the 72 sites have two- or three-coordinate (planar) oxygen and bound no volume at all, and these sites are concentrated in Al_2O_3, Ga_2O_3, Pb_2O_3, Pb_3O_4, PbO_2, SnO, SnO_2, Bi_2O_3, and Bi_4O_7. Table S6 evaluates all four measures with the same conventions as Figure S13a. Where the volume is defined, its within-host correlation with E_v is weak ($R = +0.16$ over 19 host-centered points), its host sign count is mixed (3/2), and controlling for d_{min} leaves it at a partial $R = +0.12$. The normalized volume passes the host-sign test of Figure S13a only narrowly (4 of 5 counted hosts), with an equally weak pooled correlation ($R = +0.15$). The distortion index is the strongest of the regularity measures, but it shows a Simpson-type inversion: its pooled centered correlation is negative (-0.17) while six of the nine hosts with at least three sites are individually positive, so a global coefficient would be fit to a sign opposite to that observed within most hosts; the Madelung potential and the RMS displacement of Figure S13a invert in the same way. The oxygen off-centering carries little within-host signal ($R = -0.14$) and none once d_{min} is controlled for (partial $R = -0.01$).

Table S6: Polyhedral-shell candidate descriptors evaluated against E_v with the conventions of Figure S13a: number of sites where defined, global Pearson R, pooled within-polymorph (host-centered) R with its number of centered points, per-host correlation sign count (hosts with at least three sites), and the pooled centered partial R controlling for d_{min}. The shortest metal–oxygen bond length of the vacant site, d_{min}, is listed for reference; $\langle d\rangle$ is the mean bond length of the cation shell.

descriptor	n	global R	centered R	n_{cen}	count (+/−)	partial $R \mid d_{min}$
V_{shell}	45	−0.01	+0.16	19	3/2	+0.12
$V_{shell}/\langle d\rangle^3$	45	+0.23	+0.15	19	4/1	+0.10
distortion index	72	−0.39	−0.17	47	6/3	−0.29
O off-centering	72	−0.43	−0.14	47	4/5	−0.01
d_{min} (reference)	72	−0.39	+0.65	47	9/0	—

Shortest-bond descriptor and a descriptor-library screen

The shortest metal–oxygen bond length d_{min} is the strongest host-sign-consistent within-polymorph descriptor (Figure S13a,b), yet its *raw* value correlates negatively with E_{v} across hosts: larger, softer cations give both longer bonds and lower formation energies, so the between-host trend is dominated by cation size. If the within-host trend truly reflects local bond strain, referencing d_{min} to a tabulated natural metal–oxygen bond length should remove the size scale and align the two levels. Table S7 tests five literature reference scales. Because any reference that is constant within a host cancels under host-centering, all variants except the coordination-resolved bond-valence ideal retain the within-host statistics of raw d_{min} by construction; the discriminating quantity is the global correlation and its sign.

Referencing to the bond-valence parameter R_0 (the M–O distance of unit bond valence) or to a sum of empirical (Slater-type) atomic radii[S63] aligns the global correlation with the within-host one ($-0.39 \rightarrow +0.42$ and $+0.43$, respectively) at no cost to the nine-of-nine host count: on either scale, sites compressed below the natural bond length have the lower vacancy formation energies in every host. The R_0 used here is the element-level atomic-size estimate of O'Keeffe and Brese[S61] implemented in PYMATGEN,[S42] one valence-independent value per metal, rather than the valence-resolved tabulated entries of Brese and O'Keeffe.[S64] The data do not discriminate between the ionic-calibrated and atomic references: both succeed because both remove the cation-size scale. The two remaining scales do not remove it. Cordero covalent radii[S65] track the observed bond lengths themselves, so the subtraction leaves the size trend intact (global $R = -0.38$, essentially unchanged from raw d_{min}). Shannon effective ionic radii[S66] are undefined for several of the relevant oxidation states (no tabulated radii exist for Sn^{2+} or for the intermediate valences of the mixed-valent Pb, Bi, and Tl oxides), which removes 17 of the 72 sites (three of the nine hosts with at least three sites among them; on the remaining 55 sites, raw d_{min} itself gives the same within-host $R = +0.65$ and six-of-six host count as the Shannon row), and align the global and within-host correlations only weakly where defined.

The coordination-resolved bond-valence ideal distance, $d_{\mathrm{ideal}} = R_0 - 0.31\ln(2/\mathrm{CN_O})$, converts each site's d_{min} into an absolute strain spanning -0.23 Å to $+0.23$ Å (median -0.07 Å) with the same positive sign at both levels, though its within-host correlation is diluted by the coordination-number term it introduces. The decay length 0.31 Å in d_{ideal} is that of the PYMATGEN bond-valence implementation, whereas O'Keeffe and Brese fitted the atomic parameters from which R_0 is built with 0.37 Å; with that value, the strain would span -0.29 Å to $+0.15$ Å and the within-host correlation would fall from $+0.09$ to $+0.00$, with neither sign nor the host count changed.

The pooled within-host slope $\mathrm{d}E_{\mathrm{v}}/\mathrm{d}d_{\mathrm{min}}$ is $+3.8\,\mathrm{eV\,Å^{-1}}$ ($+5.8\,\mathrm{eV\,Å^{-1}}$ for the R_0-referenced global trend), of the order of the natural force scale of a metal–oxygen bond: one bond dissociation energy per bond-valence decay length, $E_{\mathrm{b}}/b \approx$ 3–10 $\mathrm{eV\,Å^{-1}}$ for the dataset's per-bond E_{b} of 0.9–3.1 eV and $b = 0.31$ Å (the conclusion holds with the 0.37 Å decay length above). The within-host trend thus reflects local bond compression, at the force scale of the bonds themselves, not cation size.

A screen of a published descriptor library tests whether the saturation of the candidate search above reflects how those candidates were chosen. Of 127 geometric fingerprints of the vacant site from the MATMINER library[S62] (order parameters, Voronoi and bond-orientational descriptors, and radial symmetry functions), screened against the within-host residual under a false-discovery-rate control, none passes every sign-consistency and leave-one-host-out test (one polymorph held out at a time). The closest, an order parameter that is large for oxygen sites whose two markedly shorter metal–oxygen bonds subtend about 120°, survives the false-discovery-rate control only when the metastable hosts are included, and even then its sign is consistent in only six of eight hosts; in the room-temperature window, it is consistent in four of six and does not survive the control. A regularized (ridge) fit over all 127 fingerprints together with d_{min} nevertheless recovers more of the within-host variance in leave-one-host-out tests than d_{min} alone: 57 % against 39 % in the room-temperature window and 63 % against 40 % over all sites (each fingerprint standardized on the training hosts only, with the held-out values clipped to the training range). The within-host signal beyond d_{min} is therefore spread over many fingerprints rather than carried by one nameable descriptor. None of the 84 host-level composition and structure fingerprints from the same library explains the per-host offsets in the room-temperature window; the strongest of them are collinear with V_{r} and E_{g} and carry no information beyond the ionicity those two descriptors already encode.

Table S7: Reference bond-length scales for $d_{\min}$, evaluated against E_{v} over the 72 clean sites: number of sites where defined, global Pearson R, pooled within-polymorph (host-centered) R, per-host correlation sign count (hosts with at least three sites), and the median and range of the referenced $d_{\min}$, that is, $d_{\min}$ minus the reference distance (Å; a sum of tabulated metal and oxygen radii for the three radius scales). Composition-constant references cancel under host-centering and therefore share the within-host statistics of raw $d_{\min}$ over the sites where they are defined.

reference scale	n	global R	within R	count (+/−)	median	range
none (raw $d_{\min}$)	72	−0.39	+0.65	9/0	2.14	1.62 to 2.91
bond-valence $R_0 - 0.31 \ln(2/\mathrm{CN_O})$	72	+0.38	+0.09	5/4	−0.07	−0.23 to +0.23
bond-valence R_0	72	+0.42	+0.65	9/0	+0.11	−0.04 to +0.66
Shannon ionic radii (CN VI)	55	+0.22	+0.65	6/0	−0.17	−0.36 to +0.03
covalent radii (Cordero)	72	−0.38	+0.65	9/0	+0.02	−0.26 to +0.43
atomic radii (Slater-type)	72	+0.43	+0.65	9/0	−0.06	−0.36 to +0.18

Comparison with quantum-chemical bonding descriptors

Integrated crystal orbital Hamilton populations (ICOHPs) and related projected bonding indicators are widely used measures of covalent bond strength in solids, and the quantum-chemical bonding database of Naik et al.[S67] provides them precomputed, at the PBE level on Materials Project structures, for 1520 semiconductors and insulators. Fifteen training-set hosts appear in that database (17 of the 72 clean vacancy sites): the antifluorite alkali oxides and layered Cs_2O, the rock-salt alkaline-earth oxides, wurtzite BeO, the corundum-type sesquioxides of Al, Ga, and In, monoclinic Ga_2O_3, and rutile SnO_2. None of the low-symmetry, multi-site hosts that dominate the outlier analysis (the Sn, Pb, Tl, and Bi oxides other than those listed) are covered, so the comparison probes only the between-family variation. From each host's summarized bonding analysis, we extracted the ICOHP summed over the bonds at each symmetry-inequivalent oxygen site, the corresponding per-bond mean and bond count, the antibonding fraction of those bonds below the Fermi level, the oxygen Mulliken and Löwdin charges, and the two Madelung energies; every shared host has a single symmetry-inequivalent oxygen site except monoclinic Ga_2O_3, whose three sites were averaged.

Table S8 compares the host-level correlations. The oxygen-site ICOHP sum is uncorrelated with both the host-mean E_v ($R = +0.02$) and ΣE_b ($R = +0.05$), because the ICOHP registers only the covalent contribution to bonding: across the ionicity range of these oxides, In_2O_3 carries five times the oxygen-site ICOHP magnitude of BaO (-12.5 eV versus -2.5 eV) yet a 1.7 eV *lower* mean E_v, whereas the cohesive-energy route to ΣE_b captures the ionic and covalent dissociation cost together. In leave-one-out cross-validation on the 15 hosts, Huber fits to the bonding-database features reach mean absolute errors of 0.61 to 1.08 eV depending on the feature subset (1.05 eV for the integrated crystal orbital Hamilton population sum and the Mulliken charge of the oxygen site, 0.61 eV for all eight features), against a mean absolute deviation of the host-mean energies themselves of 1.27 eV, while the four crystal features reach 0.26 eV on the same hosts. The comparison carries two caveats: the database descriptors are evaluated at the PBE level on Materials Project structures rather than at r^{2}SCAN on our relaxed cells, and the shared hosts are the high-symmetry, single-oxygen-site portion of the training set.

Table S8: Host-level Pearson correlations R with the host-mean oxygen vacancy formation energy E_v on the 15 training-set hosts shared with the quantum-chemical bonding database of Naik et al.[S67] Crystal-feature descriptors are computed as in the main text; bonding descriptors are extracted from the database: site-resolved quantities at the oxygen site(s) of each host and the Madelung energy of the whole structure divided by its number of atoms. ICOHP values follow the bonding-negative sign convention.

descriptor family	descriptor	R
crystal features	ΣE_b	+0.95
	V_r	−0.94
	E_g	+0.90
	E_{hull}	−0.35
bonding database	ICOHP sum at O	+0.02
	ICOHP per bond at O	+0.13
	bond count at O	+0.03
	antibonding fraction at O	−0.28
	O Mulliken charge	−0.23
	O Löwdin charge	−0.22
	Madelung energy per atom (Mulliken)	−0.61
	Madelung energy per atom (Löwdin)	−0.68

Isolated-atom reference energies

The r^{2}SCAN-consistent route to the crystal bond dissociation energies evaluates each cohesive energy from the total energies of the isolated atoms (extended methods above), so every ΣE_b of this route depends on them (the experimental route does not); the O_2 entry additionally fixes the oxygen chemical potential μ_O of Eq. S1 and, together with the O entry, the computed O_2 dissociation energy quoted with that equation. Every species was computed spin-polarized in the same $14 \times 15 \times 16$ Å orthorhombic box (r^{2}SCAN, `ISPIN` = 2, `ISYM` = 0; with symmetry off, each atom with an open p shell breaks spherical symmetry and converges to integer orbital occupations, and the unequal box lengths only remove exact degeneracies, of about 1 meV), with the magnetic moment initialized at the Hund's-rules value m_Hund of the ground configuration for the 17 metal atoms and at $3\,\mu_\mathrm{B}$ per atom for O and O_2, and relaxed self-consistently to m_conv in every case. Table S9 therefore records, for each species, the valence configuration and ground-state term symbol from which the Hund's-rules magnetic moment follows, together with the Hund's-rules and converged moments and the total energy. Every converged moment reproduces its Hund's-rules value, confirming that no atomic reference relaxed into an excited spin state.

Table S9: Isolated-atom and O_2 reference energies used in the r^{2}SCAN-consistent cohesive energies (and, via $\mu_\mathrm{O} = \frac{1}{2}E[\mathrm{O_2}]$, the oxygen chemical potential). m_Hund is the Hund's-rules magnetic moment of the ground configuration and m_conv the self-consistently converged moment of the spin-polarized r^{2}SCAN calculation.

Species	Valence configuration	Term symbol	m_Hund (μ_B)	m_conv (μ_B)	E (eV)
Li	$2s^1$	^2S	1	1.00	−0.77665
Na	$3s^1$	^2S	1	1.00	−2.47689
K	$4s^1$	^2S	1	1.00	−5.16182
Rb	$5s^1$	^2S	1	1.00	−13.16995
Cs	$6s^1$	^2S	1	1.00	−24.40641
Be	$2s^2$	^1S	0	0.00	−0.54752
Mg	$3s^2$	^1S	0	0.00	−2.46616
Ca	$4s^2$	^1S	0	0.00	−4.96060
Sr	$5s^2$	^1S	0	0.00	−13.30753
Ba	$6s^2$	^1S	0	0.00	−24.44523
Al	$3s^2 3p^1$	$^2P^\circ$	1	1.00	−3.13609
Ga	$4s^2 4p^1$	$^2P^\circ$	1	1.00	−8.61115
In	$5s^2 5p^1$	$^2P^\circ$	1	1.00	−20.18047
Tl	$6s^2 6p^1$	$^2P^\circ$	1	1.00	−51.09820
Sn	$5s^2 5p^2$	^3P	2	2.00	−21.48913
Pb	$6s^2 6p^2$	^3P	2	2.00	−53.46650
Bi	$6s^2 6p^3$	$^4S^\circ$	3	3.00	−56.19935
O	$2s^2 2p^4$	^3P	2	2.00	−3.21718
O_2	—	$^3\Sigma_g^-$	2	2.00	−11.86096

Additional supporting data

The archive `cfm_si_data.zip` accompanies this Supporting Information and holds the machine-readable data behind the main-text figures and behind the tables and figures of this document. Its top-level directory `cfm_si_data/` contains `README.txt` (purpose, conventions, third-party attribution, and file list; TXT), `COLUMNS.txt` (column dictionary; TXT), `TRANSFORMS.txt` (path rewrites applied at packaging; TXT), `MANIFEST.sha256` (SHA-256 checksum of every file; TXT), and seven subdirectories, whose files are listed below with a brief description and the file type of each:

- `model/`: the fitted crystal-feature-model coefficients and the cross-validation statistics for the two-descriptor model and the full four-descriptor form (dense-matched reference, r^{2}SCAN fundamental gap) and for each reference and band-gap variant discussed in the Results, with per-site predictions of the primary model and of the two- and four-descriptor all-site models.
  - `cfm_fit_results.json`: fitted coefficients, intercepts, and cross-validation statistics over 1000 random splits for the primary two-descriptor model and 13 variant forms (JSON).
  - `cfm_predictions.csv`: per-site consensus predictions and residuals of three models, the primary model and the two- and four-descriptor all-site models, for the 72 clean vacancy sites (CSV).
  - `descriptor_ablation.txt`: descriptor-ablation ladder with the test MAE, R^2, and native coefficients of five nested model forms on both site subsets (TXT).
  - `grouped_cv.txt`: leave-one-polymorph-out and leave-one-composition-out MAEs of the four- and two-descriptor models on both site subsets (TXT).
  - `grouped_cv.json`: the same grouped cross-validation record in machine-readable form, with the per-composition errors of the primary model printed in Table S5 (JSON).
  - `cfm_fit_variants_sitevr.json`: paired-fold refits replacing the composition-level descriptors by site-resolved variants, with out-of-fold predictions (JSON).
  - `signed_stability_analysis.txt`: variance, single-descriptor correlations, and collinearity of the descriptors, with cross-validated refits that replace E_{hull} by the signed reduction energy (TXT).
  - `signed_stability_analysis.json`: the same analysis in machine-readable form, the source of the signed-stability correlations, slope, and test MAEs quoted in the main text and in the model-robustness section above (JSON).
  - `ladder/ladder_paired.json`: four-descriptor refits of all five reference-ladder rungs on one set of paired splits, with the paired step differences, the values plotted in Figure S11 (JSON).
  - `ladder/cfm_fit_results_perfectGamma_KMISMATCH.json`: four-descriptor fit statistics of the first reference-ladder rung of Figure S11, the Γ-only perfect-supercell reference (JSON).
  - `ladder/cfm_fit_results_multEprim.json`: four-descriptor fit statistics of the second rung, the multiplicity-scaled bulk-cell reference (JSON).
  - `ladder/cfm_fit_results_matchedk_mpgap.json`: four-descriptor fit statistics of the third rung, the matched-cell, matched-k perfect-supercell reference with the Materials Project generalized-gradient gap (JSON).
  - `ladder/cfm_fit_results_matchedkGamma.json`: fit statistics of the fourth rung, the matched-k reference with the r^{2}SCAN fundamental gap on the defect supercells before their dense re-relaxation (JSON).
- `descriptors/`: the tabulated descriptor values (crystal bond dissociation energy, crystal reduction potential, band gap, and energy above the convex hull) for all 72 clean vacancy sites, the first two by both the r^{2}SCAN-consistent and the experimental-thermochemistry routes (the band gap and hull energy have the r^{2}SCAN route only), with the reference energies and thermochemical inputs from which they were built.
  - `cfm_fit_table.csv`: per-site formation energies and the four descriptors, the bond dissociation energy and reduction potential by both routes, for all 72 sites of the fit, with the clean flag (CSV).

  - `composition_descriptors_dft.csv`: per-composition cohesive energy, bond count, per-bond E_b, and adjacent-couple V_r at the r^{2}SCAN level (CSV).
  - `composition_descriptors_exp.csv`: per-composition experimental-thermochemistry E_b and V_r with reliability flags (CSV).
  - `vr_reductions_dft.json`: every accessible reduction couple per composition with its r^{2}SCAN reaction energy and potential (JSON).
  - `experimental_thermo.json`: tabulated formation enthalpies and cohesive energies with their sources (JSON).
  - `r2scan_gaps.csv`: r^{2}SCAN fundamental, direct, Γ-point, and density-of-states gaps per polymorph stage from the bulk relaxation, with the Materials Project generalized-gradient gap (CSV).
  - `r2scan_gaps_firmedup.csv`: dense-mesh static band gaps of 44 polymorph stages, the source of E_g for every polymorph in the fit, with the metallic hosts set to zero (CSV).
  - `r2scan_ehull.csv`: per-polymorph r^{2}SCAN formation enthalpy and energy above the convex hull, with the Materials Project value and the pre-correction first-pass record (CSV).
  - `reference_energies.json`: O_2 and O reference energies, μ_O, and the O_2 dissociation energy (JSON).
  - `atom_energies.csv`: isolated-atom and O/O_2 reference energies and converged moments (CSV).
  - `metal_solid_energies.csv`: elemental-metal reference energies per atom from the k-converged statics, with the earlier spin-polarized values, the phase rule, and Materials Project identifiers (CSV).

- `sweep/`: the energy-above-hull threshold sweep summarized in the Results, which quantifies the sensitivity of the fit to the room-temperature-stability window.
  - `ehull_threshold_sweep.json`: fit statistics and coefficients at four E_hull cutoffs for both model forms (JSON).

- `bonding_screen/`: the host-level bonding-descriptor screen underlying Table S8, tabulating the per-host crystal-feature and bonding-database descriptor values with the correlation and cross-validation statistics.
  - `naik_bonding_screen.csv`: host-level crystal-feature and bonding-database descriptors with the host-mean E_v for the 15 shared hosts (CSV).
  - `naik_bonding_screen.json`: correlation, cross-correlation, leave-one-out MAE, and provenance summary of the screen (JSON).
  - `naik_host_mpids.csv`: mapping of the 15 retained polymorphs shared with the bonding database to their Materials Project identifiers (CSV).

- `site_features/`: the per-site candidate featurizations and defect-level characteristics underlying Figure S13, including the polyhedral-shell measures of Table S6 and the reference-bond-length analysis of Table S7; the supercells on which each quantity was evaluated are given in the Evaluation conventions subsection below.
  - `site_weighted_descriptors.csv`: shell geometry (d_min and mean bond length), fractional coordination, bond-valence sum, and bond-weighted descriptor variants for the 72 clean sites (CSV).
  - `site_vr_valence.csv`: oxidation-state-resolved cation-shell reduction potentials and bond sums for the 15 sites of the four mixed-valence compositions (CSV).
  - `site_madelung_dense.csv`: point-ion Madelung potential at the vacant site and relaxation magnitudes of the ground-state defect supercell for the 72 clean sites (CSV).
  - `site_o2p_centers_dense.csv`: occupied O-2p band centers of the removed oxygen from the dense-matched pristine supercells (CSV).

  - `site_finalstate_dense.csv`: Kohn–Sham donor depth, defect-level localization (inverse participation ratio), orbital character, and site moments of each ground-state defect supercell (CSV).
  - `site_polyhedral.csv`: polyhedral-shell volume, normalized volume, distortion index, and oxygen off-centering per clean site (CSV).
  - `polyhedral_summary.json`: host-centered correlation statistics of the polyhedral-shell measures, the source of Table S6 (JSON).
  - `dmin_reference_scales.csv`: shortest metal–oxygen bond per site referenced to the bond-valence, Shannon, covalent, and atomic length scales (CSV).
  - `dmin_reference_summary.json`: host-centered correlation statistics of the referenced shortest bonds, the source of Table S7 (JSON).
  - `matminer_screen_table.csv`: within-host and host-level screen statistics of the 127 site and 84 host MATMINER fingerprints (CSV).

- `si_tables/`: the source data of the tables and figures of this document.
  - `si_kgrid.csv`: supercell size, multiplicity, and production k-mesh per fit-set polymorph stage, as printed in Table S1 (CSV).
  - `symmetry_audit.csv`: space-group reproduction audit of every polymorph with per-stage final energies, symmetry-tolerance scans, and verdicts, the source of Table S2 and Figures S3 to S5 (CSV).
  - `symmetry_audit_SI.csv`: Table S2 as printed (CSV).
  - `refcheck_basin.csv`: per-polymorph outcome of the rattled re-relaxation of each pristine reference supercell, with the energy change, structure match, atomic displacements, and shortest O–O distances, the source of Table S3 (CSV).
  - `si_polymorphs.csv`: Table S4 as printed (CSV).
  - `mixedvalence_stats.json`: bond-valence-sum oxidation-state statistics of the four mixed-valence hosts of Figure S7 (JSON).
  - `sno_collapse_stats.json`: coordination statistics of the red-SnO host collapse of Figure S8 (JSON).
  - `pbo2_collapse_stats.json`: coordination and relaxation statistics of the cubic-PbO_2 vacancy collapse of Figure S9 (JSON).

- `structures/`: the relaxed structures and the absolute total energies of the calculations that give the vacancy formation energies, the descriptors, and the convex hull of the main text and the energy differences of Table S2; the structures behind the other checks of this document are part of the NOMAD dataset described below. Each energy is the VASP total energy extrapolated to zero smearing at the final geometry, as read by the analysis; no vibrational-frequency calculations were performed.
  - `structures_energies.csv`: one row for each structure file and for each of the 18 isolated-atom references (285 rows), with the role of the structure, its composition and number of atoms, the k-point mesh, the total energy in eV and in hartree, and the source calculation in the NOMAD dataset (CSV).
  - `*.vasp`: the 267 relaxed structures, one file each in VASP POSCAR format with Cartesian coordinates: the higher-symmetry and rattled bulk cells of all 66 polymorphs (132 files), the pristine reference supercell of each of the 39 retained polymorphs, the ground-state vacancy supercell of each of the 72 clean sites, the O_2 molecule, the 17 elemental-metal reference cells, and both cells of the three phases that enter only the convex hull (POSCAR).

Figure S7 renders the deposited pristine relaxations. In Figures S8 and S9, panel (a) renders the deposited Γ-point pristine supercells and panel (b) a screening-stage defect calculation, deposited with the SHAKENBREAK configuration record for Figure S9 but not for Figure S8; the archive holds the statistics quoted in all three captions.

The tables of the archive, together with the intermediate tables from which they are assembled, the analysis scripts that regenerate every table and figure, the figure files, and the software-environment specifications, will be deposited in a Zenodo record on publication, under the CC BY 4.0 license for the data and figures and the MIT license for the scripts. The production r^{2}SCAN calculations will be deposited in the NOMAD repository as a dataset on publication. The dataset holds the VASP input and output files of every calculation, except the `POTCAR` files, which are distributed under the VASP license, and the wavefunction (`WAVECAR`) files; the potentials are those listed in the extended methods. The calculations comprise the pristine relaxations of all 66 polymorphs, the dense-matched pristine and defect supercells, the dense-mesh band-gap statics, the rattled re-relaxations of the pristine reference supercells, the perfect-supercell, elemental-metal, isolated-atom, and O_2 reference calculations with the convergence, spin-state, fresh-basis, and suboxide calculations behind the reference energies, and a record of the ShakeNBreak configuration search (the inputs, starting and final structures, and final energy of every configuration of every fitted site and of the high-pressure polymorphs, with the full output of each site's ground state). The dataset README records the fourteen truncated `vasprun.xml` outputs that were removed (their `OUTCAR` files remain) and the twelve aborted runs that were dropped (four symmetry-tolerance retries and eight static calculations that stopped at startup).

Evaluation conventions for the electrostatic, bond-valence, band-center, and defect-level quantities

The defect-level characteristics and the O-2p band centers of `site_features/` were evaluated on the dense-matched production supercells described in the main-text Methods; the initial-state geometry measures (the shell geometry, fractional coordination, and bond-valence sums; the polyhedral-shell measures; the site-resolved reduction potentials of the mixed-valence hosts; and the Madelung potentials and relaxation magnitudes) were evaluated on the tiled pristine supercells and ground-state defect geometries of the screening stage from which the production supercells were built. Re-evaluating the shell-geometry, bond-valence, and polyhedral measures on the production supercells changes their pooled host-centered correlations with E_v by up to 0.008.

The initial-state candidates of `site_features/` follow fixed conventions: the point-ion Madelung potential at the vacant site is the electrostatic potential of the formal-charge lattice (oxygen as -2, cations at their nominal oxidation state, with the bond-valence-split states of Figure S7 for the mixed-valence hosts), evaluated by Ewald summation at the PYMATGEN default accuracy on the pristine relaxed supercell; the bond-valence sum of the removed oxygen is $\sum_j \exp[(R_0 - d_j)/b]$ over the cations within 4.0 Å, the same shell within which d_min is taken (any larger cutoff changes the sum by less than 0.5 %), with $b = 0.31$ Å, the decay length of the PYMATGEN bond-valence implementation; and the O-2p band center is the centroid of the occupied site-projected O-2p density of states ($E \leq E_\mathrm{F}$) of the dense-matched pristine supercell, referenced to the Fermi level of that calculation (placed at midgap in the gapped hosts by the VASP setting `EFERMI = Midgap`), a reference that cancels in the within-host comparisons of Figure S13a.

The defect-level characteristics are extracted from each site's ground-state defect supercell: the defect level is taken as the highest occupied Kohn–Sham level (occupation threshold 0.5, both spin channels, all k-points; the occupied bands within 0.02 eV of the highest occupied eigenvalue, in the spin channel and at the k-point where it is attained, are counted as degenerate with the level), the donor depth $\varepsilon_\mathrm{LU} - \varepsilon_\mathrm{HO}$ is the separation between the lowest unoccupied and highest occupied eigenvalues of that supercell, and the localization and orbital character of the defect level follow from the site- and orbital-projected weights (`LORBIT = 11`, read from the full-precision `vasprun.xml` output; `PROCAR` rounds them to three decimals) summed over the level and the bands degenerate with it. The localization is reported as the inverse participation ratio (IPR) of the per-atom weights w_i, $\sum_i w_i^2/(\sum_i w_i)^2$, whose reciprocal is the effective number of participating atoms, a median of 13 across the 72 sites (the defect-level IPR of Figure S13a), and the $s/p/d$ character as fractions of the metal-site weight. The site magnetic moments quoted in the descriptor-search section above are the site-projected totals of the final magnetization block of the VASP output; they omit the magnetization outside the atomic projection spheres, so their sum can fall short of the net moment of the cell (BaO, site 8: 0.25 against 1.06 μ_B). The relaxation magnitude is the root-mean-square (RMS) displacement of the atoms of the ShakeNBreak ground-state configuration, each matched to the nearest pristine atom of the same species under the minimum-image convention.

References

(S1) Wexler, R. B.; Sai Gautam, G.; Stechel, E. B.; Carter, E. A. Factors Governing Oxygen Vacancy Formation in Oxide Perovskites. *J. Am. Chem. Soc.* **2021**, *143*, 13212–13227.

(S2) Zagorac, D.; Müller, H.; Ruehl, S.; Zagorac, J.; Rehme, S. Recent Developments in the Inorganic Crystal Structure Database: Theoretical Crystal Structure Data and Related Features. *J. Appl. Crystallogr.* **2019**, *52*, 918–925.

(S3) Mori, Y.; Niiya, N.; Ukegawa, K.; Mizuno, T.; Takarabe, K.; Ruoff, A. L. High-Pressure X-Ray Structural Study of BeO and ZnO to 200 GPa. *Phys. Status Solidi B* **2004**, *241*, 3198–3202.

(S4) Liu, L.-G. A Dense Modification of BaO and Its Crystal Structure. *J. Appl. Phys.* **1971**, *42*, 3702–3704.

(S5) Weir, S. T.; Vohra, Y. K.; Ruoff, A. L. High-Pressure Phase Transitions and the Equations of State of BaS and BaO. *Phys. Rev. B* **1986**, *33*, 4221–4226.

(S6) Lin, J.-F.; Degtyareva, O.; Prewitt, C. T.; Dera, P.; Sata, N.; Gregoryanz, E.; Mao, H.-K.; Hemley, R. J. Crystal Structure of a High-Pressure/High-Temperature Phase of Alumina by In Situ X-Ray Diffraction. *Nat. Mater.* **2004**, *3*, 389–393.

(S7) Dinnebier, R. E.; Carlson, S.; Hanfland, M.; Jansen, M. Bulk Moduli and High-Pressure Crystal Structures of Minium, Pb_3O_4, Determined by X-Ray Powder Diffraction. *Am. Mineral.* **2003**, *88*, 996–1002.

(S8) Grocholski, B.; Shim, S.-H.; Cottrell, E.; Prakapenka, V. B. Crystal Structure and Compressibility of Lead Dioxide up to 140 GPa. *Am. Mineral.* **2014**, *99*, 170–177.

(S9) Serebryanaya, N. R.; Kabalkina, S. S.; Vereshchagin, L. F. The Crystal Structure of Tin Oxide at High Pressure. *Dokl. Akad. Nauk SSSR* **1969**, *187*, 307–310.

(S10) Haines, J.; Léger, J. M. X-Ray Diffraction Study of the Phase Transitions and Structural Evolution of Tin Dioxide at High Pressure: Relationships between Structure Types and Implications for Other Rutile-Type Dioxides. *Phys. Rev. B* **1997**, *55*, 11144–11154.

(S11) Locherer, T.; Prasad, D. L. V. K.; Dinnebier, R.; Wedig, U.; Jansen, M.; Garbarino, G.; Hansen, T. High-Pressure Structural Evolution of HP-Bi_2O_3. *Phys. Rev. B* **2011**, *83*, 214102.

(S12) Gerlach, W. Die Gitterstruktur der Erdalkalioxyde. *Z. Phys.* **1922**, *9*, 184–192.

(S13) McKeehan, L. W. The Crystal Structure of Beryllium and of Beryllium Oxide. *Proc. Natl. Acad. Sci. U. S. A.* **1922**, *8*, 270–274.

(S14) Aminoff, G. Über Berylliumoxyd als Mineral und dessen Kristallstruktur. *Z. Kristallogr.* **1925**, *62*, 113–122.

(S15) Zachariasen, W. H. Über die Kristallstruktur von BeO. *Nor. Geol. Tidsskr.* **1925**, *8*, 189–200.

(S16) Alptekin, S.; Durandurdu, M. Pressure-Induced Phase Transition of BeO. *Solid State Commun.* **2009**, *149*, 345–348.

(S17) Christoph, G. G.; Corbató, C. E.; Hofmann, D. A.; Tettenhorst, R. T. The Crystal Structure of Boehmite. *Clays Clay Miner.* **1979**, *27*, 81–86.

(S18) Ewing, F. J. The Crystal Structure of Diaspore. *J. Chem. Phys.* **1935**, *3*, 203–207.

(S19) Vitse, P.; Galy, J.; Potier, A. Obtention, structure et composition d'un oxyhydroxyde de gallium III : GaOOH. *C. R. Acad. Sci., Ser. C* **1973**, *277*, 159–162.

(S20) Isobe, M.; Onoda, M.; Shizuya, M.; Tanaka, M.; Takayama-Muromachi, E. Modulated Structure of the Composite Crystal $[2CaOH]_kCoO_2$ ($k = 0.576$). *J. Phys. Soc. Jpn.* **2007**, *76*, 014602.

(S21) Latrach, A.; Mentzen, B. F.; Bouix, J. Polymorphisme du dioxotriplomb(II)tétraoxosulfate(VI) $Pb_3O_2(SO_4)$ à basse température. II – Étude structurale de la phase ferroélastique γ-$Pb_3O_2(SO_4)$ par diffraction neutronique sur poudres à 2 et 45 K. *Mater. Res. Bull.* **1985**, *20*, 1081–1090.

(S22) Matsumoto, A.; Koyama, Y.; Tanaka, I. Structures and Energetics of Bi_2O_3 Polymorphs in a Defective Fluorite Family Derived by Systematic First-Principles Lattice Dynamics Calculations. *Phys. Rev. B* **2010**, *81*, 094117.

(S23) Harwig, H. A. On the Structure of Bismuthsesquioxide: The α, β, γ, and δ-Phase. *Z. Anorg. Allg. Chem.* **1978**, *444*, 151–166.

(S24) Gattow, G.; Schütze, D. Über Wismutoxide. VI. Über ein Wismut(III)-oxid mit höherem Sauerstoffgehalt (β-Modifikation). *Z. Anorg. Allg. Chem.* **1964**, *328*, 44–68.

(S25) Johnsen, R. E.; Norby, P. A Structural Study of Stacking Disorder in the Decomposition Oxide of MgAl Layered Double Hydroxide: A DIFFaX+ Analysis. *J. Phys. Chem. C* **2009**, *113*, 19061–19066.

(S26) Zav'yalova, A. A.; Imamov, R. M. Special Features of the Crystal Structure of Bismuth Oxides. *J. Struct. Chem.* **1973**, *13*, 811–814, Translated from Zh. Strukt. Khim. **1972**, *13*, 869–873.

(S27) Köhler, J.; Tong, J.; Dinnebier, R.; Simon, A. Crystal Structure and Electronic Structure of Red SnO. *Z. Anorg. Allg. Chem.* **2012**, *638*, 1970–1975.

(S28) Wittmann, S.; Murshed, M. M.; Gesing, T. M. On Red Tin(II) Oxide: Temperature-Dependent Structural, Spectroscopic, and Thermogravimetric Properties. *Z. Anorg. Allg. Chem.* **2022**, *648*, e202200311.

(S29) Reuter, H. Red and Blue-Black Tin Monoxide, SnO: Pitfalls, Challenges, and Helpful Tools in Crystal Structure Determination of Low-Intensity Datasets from Microcrystals. *Crystals* **2023**, *13*, 1281.

(S30) Atou, T.; Faqir, H.; Kikuchi, M.; Chiba, H.; Syono, Y. A New High-Pressure Phase of Bismuth Oxide. *Mater. Res. Bull.* **1998**, *33*, 289–292.

(S31) Hill, R. J. Refinement of the Structure of Orthorhombic PbO (Massicot) by Rietveld Analysis of Neutron Powder Diffraction Data. *Acta Crystallogr., Sect. C: Cryst. Struct. Commun.* **1985**, *41*, 1281–1284.

(S32) Sabrowsky, H. Zur Darstellung und Kristallstruktur von Tl_2O. *Z. Anorg. Allg. Chem.* **1971**, *381*, 266–279.

(S33) Smith, D. K.; Cline, C. F.; Austerman, S. B. The Crystal Structure of β-Beryllia. *Acta Crystallogr.* **1965**, *18*, 393–397.

(S34) Kresse, G.; Furthmüller, J. Efficient Iterative Schemes for *Ab Initio* Total-Energy Calculations Using a Plane-Wave Basis Set. *Phys. Rev. B* **1996**, *54*, 11169–11186.

(S35) Blöchl, P. E. Projector Augmented-Wave Method. *Phys. Rev. B* **1994**, *50*, 17953–17979.

(S36) Kresse, G.; Joubert, D. From Ultrasoft Pseudopotentials to the Projector Augmented-Wave Method. *Phys. Rev. B* **1999**, *59*, 1758–1775.

(S37) Furness, J. W.; Kaplan, A. D.; Ning, J.; Perdew, J. P.; Sun, J. Accurate and Numerically Efficient r^2SCAN Meta-Generalized Gradient Approximation. *J. Phys. Chem. Lett.* **2020**, *11*, 8208–8215.

(S38) Furness, J. W.; Kaplan, A. D.; Ning, J.; Perdew, J. P.; Sun, J. Correction to "Accurate and Numerically Efficient r^2SCAN Meta-Generalized Gradient Approximation". *J. Phys. Chem. Lett.* **2020**, *11*, 9248.

(S39) Jain, A.; Ong, S. P.; Hautier, G.; Chen, W.; Richards, W. D.; Dacek, S.; Cholia, S.; Gunter, D.; Skinner, D.; Ceder, G.; Persson, K. A. Commentary: The Materials Project: A Materials Genome Approach to Accelerating Materials Innovation. *APL Mater.* **2013**, *1*, 011002.

(S40) Kingsbury, R.; Gupta, A. S.; Bartel, C. J.; Munro, J. M.; Dwaraknath, S.; Horton, M.; Persson, K. A. Performance Comparison of r^{2}SCAN and SCAN metaGGA Density Functionals for Solid Materials via an Automated, High-Throughput Computational Workflow. *Phys. Rev. Mater.* **2022**, *6*, 013801.

(S41) Kingsbury, R. S.; Rosen, A. S.; Gupta, A. S.; Munro, J. M.; Ong, S. P.; Jain, A.; Dwaraknath, S.; Horton, M. K.; Persson, K. A. A Flexible and Scalable Scheme for Mixing Computed Formation Energies from Different Levels of Theory. *npj Comput. Mater.* **2022**, *8*, 195.

(S42) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A Robust, Open-Source Python Library for Materials Analysis. *Comput. Mater. Sci.* **2013**, *68*, 314–319.

(S43) Perdew, J. P.; Burke, K.; Ernzerhof, M. Generalized Gradient Approximation Made Simple. *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.

(S44) Legrain, F.; Manzhos, S. Understanding the Difference in Cohesive Energies between Alpha and Beta Tin in DFT Calculations. *AIP Adv.* **2016**, *6*, 045116.

(S45) Haxhijaj, A.; Riemelmoser, S.; Pasquarello, A. Lattice Dynamics and Structural Phase Stability of Group-IV Elemental Solids with the r^{2}SCAN Functional. *Phys. Rev. B* **2026**, *113*, 104105.

(S46) Mehl, M. J.; Ronquillo, M.; Hicks, D.; Esters, M.; Oses, C.; Friedrich, R.; Smolyanyuk, A.; Gossett, E.; Finkenstadt, D.; Curtarolo, S. Tin-Pest Problem as a Test of Density Functionals Using High-Throughput Calculations. *Phys. Rev. Mater.* **2021**, *5*, 083608.

(S47) Shen, J.-X.; Voss, L. F.; Varley, J. B. Simulating Charged Defects at Database Scale. *J. Appl. Phys.* **2024**, *135*, 145102.

(S48) Mosquera-Lois, I.; Kavanagh, S. R.; Walsh, A.; Scanlon, D. O. ShakeNBreak: Navigating the Defect Configurational Landscape. *J. Open Source Softw.* **2022**, *7*, 4817.

(S49) Mosquera-Lois, I.; Kavanagh, S. R.; Walsh, A.; Scanlon, D. O. Identifying the Ground State Structures of Point Defects in Solids. *npj Comput. Mater.* **2023**, *9*, 25.

(S50) Kumagai, Y.; Tsunoda, N.; Takahashi, A.; Oba, F. Insights into Oxygen Vacancies from High-Throughput First-Principles Calculations. *Phys. Rev. Mater.* **2021**, *5*, 123803.

(S51) Witman, M. D.; Goyal, A.; Ogitsu, T.; McDaniel, A. H.; Lany, S. Defect Graph Neural Networks for Materials Discovery in High-Temperature Clean-Energy Applications. *Nat. Comput. Sci.* **2023**, *3*, 675–686.

(S52) Baldassarri, B.; He, J.; Gopakumar, A.; Griesemer, S.; Salgado-Casanova, A. J. A.; Liu, T.-C.; Torrisi, S. B.; Wolverton, C. Oxygen Vacancy Formation Energy in Metal Oxides: High-Throughput Computational Studies and Machine-Learning Predictions. *Chem. Mater.* **2023**, *35*, 10619–10634.

(S53) Freysoldt, C.; Grabowski, B.; Hickel, T.; Neugebauer, J.; Kresse, G.; Janotti, A.; Van de Walle, C. G. First-Principles Calculations for Point Defects in Solids. *Rev. Mod. Phys.* **2014**, *86*, 253–305.

(S54) Linstrom, P. J.; Mallard, W. G. NIST Chemistry WebBook, NIST Standard Reference Database Number 69: Oxygen, Constants of Diatomic Molecules. 2026; `https://webbook.nist.gov/cgi/cbook.cgi?ID=C7782447&Mask=1000`.

(S55) Chase, M. W., Jr. NIST-JANAF Thermochemical Tables, Table O-001: Oxygen, Monatomic (O). 1998; `https://janaf.nist.gov/tables/O-001.html`.

(S56) Pan, H.; Ganose, A. M.; Horton, M.; Aykol, M.; Persson, K. A.; Zimmermann, N. E. R.; Jain, A. Benchmarking Coordination Number Prediction Algorithms on Inorganic Crystal Structures. *Inorg. Chem.* **2021**, *60*, 1590–1603.

(S57) Barin, I. *Thermochemical Data of Pure Substances*, 3rd ed.; VCH: Weinheim, 1995.

(S58) Huber, P. J. Robust Estimation of a Location Parameter. *Ann. Math. Stat.* **1964**, *35*, 73–101.

(S59) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(S60) Malmros, G. The Crystal Structure of α-Bi_2O_3. *Acta Chem. Scand.* **1970**, *24*, 384–396.

(S61) O'Keeffe, M.; Brese, N. E. Atom Sizes and Bond Lengths in Molecules and Crystals. *J. Am. Chem. Soc.* **1991**, *113*, 3226–3229.

(S62) Ward, L. et al. Matminer: An Open Source Toolkit for Materials Data Mining. *Comput. Mater. Sci.* **2018**, *152*, 60–69.

(S63) Slater, J. C. Atomic Radii in Crystals. *J. Chem. Phys.* **1964**, *41*, 3199–3204.

(S64) Brese, N. E.; O'Keeffe, M. Bond-Valence Parameters for Solids. *Acta Crystallogr., Sect. B: Struct. Sci.* **1991**, *47*, 192–197.

(S65) Cordero, B.; Gómez, V.; Platero-Prats, A. E.; Revés, M.; Echeverría, J.; Cremades, E.; Barragán, F.; Alvarez, S. Covalent Radii Revisited. *Dalton Trans.* **2008**, 2832–2838.

(S66) Shannon, R. D. Revised Effective Ionic Radii and Systematic Studies of Interatomic Distances in Halides and Chalcogenides. *Acta Crystallogr., Sect. A* **1976**, *32*, 751–767.

(S67) Naik, A. A.; Ertural, C.; Dhamrait, N.; Benner, P.; George, J. A Quantum-Chemical Bonding Database for Solid-State Materials. *Sci. Data* **2023**, *10*, 610.