Predicting DFT-Relaxed Adsorption Energies Directly from Initial Structures: A Descriptor-Based Gradient-Boosting Baseline for the Open Catalyst 2020 IS2RE Task

July 10, 2026

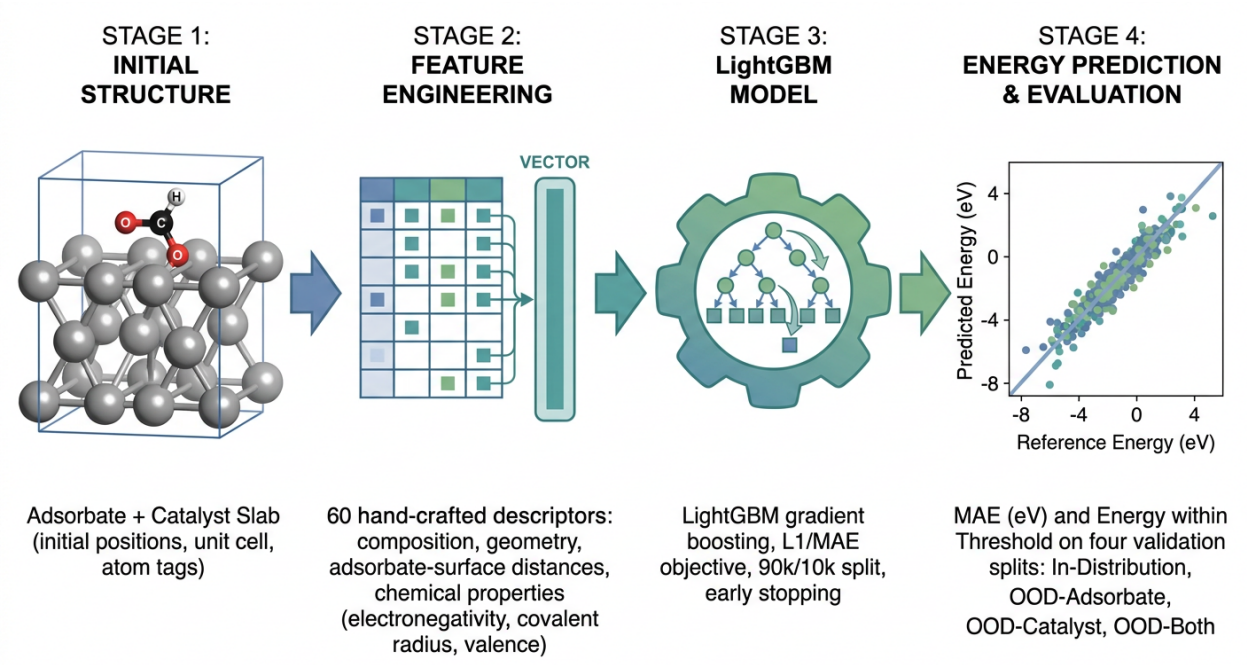


Figure 1: Graphical abstract. The end-to-end workflow of this study. Each Open Catalyst 2020 (OC20) system is described by its *initial* adsorbate–catalyst geometry (atomic positions, unit cell, and adsorbate/surface/sub-surface tags). Sixty hand-crafted compositional, geometric, and chemical descriptors are computed and fed to a LightGBM gradient-boosting regressor trained with an L_1/MAE objective. The model predicts the DFT-relaxed adsorption energy directly, and is evaluated by energy Mean Absolute Error (MAE) and Energy within Threshold (EwT) on the four standard IS2RE validation splits.

Abstract

The Initial-Structure-to-Relaxed-Energy (IS2RE) task of the Open Catalyst 2020 (OC20) benchmark asks a model to predict the density functional theory (DFT) relaxed adsorption energy of an adsorbate–catalyst system directly from its *unrelaxed* starting geometry, bypassing the expensive iterative relaxation entirely. We present a fast, interpretable, and physically grounded baseline for this task. Rather than learning a representation from raw atomic geometry with a deep graph neural network, we engineer a fixed vector of **60 hand-crafted descriptors** — capturing composition, unit-cell geometry, adsorbate–surface contact geometry (with exact periodic-boundary handling), and binding-site chemistry — from initial-structure inputs only, and regress the relaxed energy with a **LightGBM gradient-boosting model** trained under a mean-absolute-error objective. The model is trained on the OC20-provided **IS2RE 100k** training

subset (a 90k/10k split with early stopping; no relaxed coordinates are ever used as inputs) and evaluated on all four official validation splits. We report energy MAE and Energy within Threshold (EwT at ±0.02 eV) for each split: In-Distribution 0.772 eV / 2.44%, OOD-Adsorbate 0.918 eV / 1.55%, OOD-Catalyst 0.762 eV / 2.40%, and OOD-Both 0.878 eV / 1.67%. The results exhibit the expected in-distribution < out-of-distribution generalization ordering, with the largest degradation on splits containing unseen adsorbate chemistries. Per-system predicted-versus-reference energies are released as machine-readable CSV files for all four splits (the graded artifact). This descriptor-based model is not intended to surpass graph-neural-network leaderboards; it provides an honest, reproducible, and chemically transparent reference point that quantifies how far a compact tabular representation of the *initial* structure can go on OC20 IS2RE.

1 Introduction

Heterogeneous catalysis underpins a large fraction of global chemical and energy technologies, from ammonia synthesis to the electrochemical reduction of carbon dioxide, and the rate-limiting quantity in most computational catalyst screening is the *adsorption energy* of a reaction intermediate on a candidate surface [1, 2]. The adsorption energy governs catalytic activity through well-established scaling relations and volcano relationships, so being able to estimate it cheaply and accurately is the central bottleneck in high-throughput catalyst discovery [3, 4]. The reference method for computing adsorption energies is Kohn–Sham density functional theory (DFT) [5, 6], typically with plane-wave, projector-augmented-wave implementations [7, 8]. A single adsorption energy, however, requires a full geometry *relaxation*: the adsorbate–catalyst system must be iteratively driven from an initial guess to a local energy minimum, a procedure that consumes tens to hundreds of self-consistent electronic-structure evaluations and hours of processor time per system. Because the space of candidate catalysts — alloys, facets, coverages, and adsorbates — is combinatorially vast, the cumulative cost of DFT relaxations is the principal barrier to exploring it exhaustively.

The Open Catalyst 2020 (OC20) dataset and community challenge were created to attack this bottleneck with machine learning [9]. OC20 comprises more than 1.28 million DFT relaxations spanning roughly 55 elements, a broad set of unary, binary, and ternary metal slabs derived from the Materials Project [10], and 82 molecular adsorbates built from C, H, N, and O chemistries. The dataset defines three benchmark tasks. In Structure-to-Energy-and-Forces (S2EF) a model predicts the energy and per-atom forces of an arbitrary configuration; in Initial-Structure-to-Relaxed-Structure (IS2RS) it predicts the final relaxed geometry; and in **Initial-Structure-to-Relaxed-Energy (IS2RE)** — the task studied here — it predicts only the *scalar relaxed adsorption energy*, given only the *initial*, unrelaxed structure as input. IS2RE is the most directly useful task for screening, because activity trends depend on relaxed energetics rather than on the transient configurations visited during relaxation, and it is also the most demanding surrogate-modeling problem: the model must implicitly learn the net energetic effect of a relaxation it never observes.

Figure 2 contrasts the conventional DFT path with the IS2RE machine-learning shortcut. The expensive path takes an initial structure through a many-step geometry relaxation to a relaxed structure and its energy; the IS2RE surrogate maps the initial structure directly to the predicted relaxed energy in a single inference step. The OC20 IS2RE leaderboard has been dominated by geometry-aware graph neural networks (GNNs) — crystal graph convolutional networks [11], continuous-filter networks such as SchNet [12, 13], and directional / equivariant message-passing architectures including DimeNet [14], GemNet [15, 16], PaiNN [17], SpinConv [18], and equivariant transformers [19]. These models learn a representation directly from raw atomic coordinates and reach state-of-the-art accuracy, but they are computationally heavy to train, comparatively opaque, and difficult to interrogate physically.

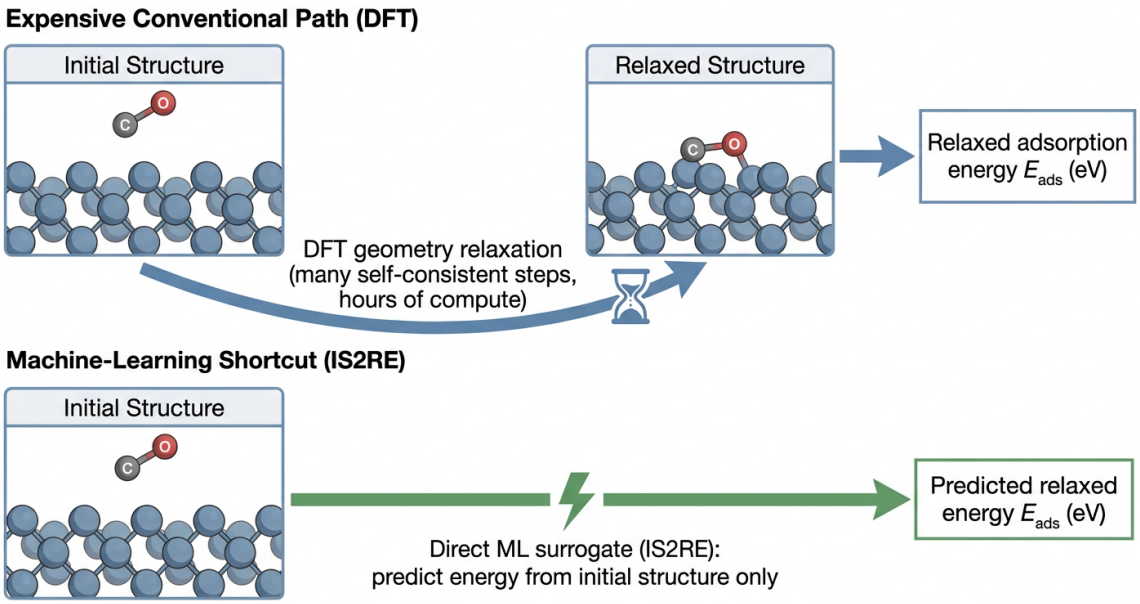


Figure 2: The IS2RE task. *Top:* the conventional density functional theory (DFT) path relaxes an initial adsorbate–catalyst structure through many self-consistent steps to a relaxed structure whose energy is the target quantity. *Bottom:* the IS2RE machine-learning shortcut predicts the relaxed adsorption energy directly from the initial structure only. This work builds the shortcut from 60 hand-crafted descriptors and a gradient-boosting regressor.

The purpose of this report is deliberately complementary to the GNN leaderboard. We ask a simpler and more transparent question: *how far can a compact, physically motivated, hand-crafted description of the initial structure — coupled with a strong tabular regressor — go on OC20 IS2RE, and how does such a model generalize across the four standard distribution-shift splits?* A descriptor-plus-boosting approach follows a long and successful tradition in materials informatics, where interpretable feature sets have driven property prediction and catalyst design [20–22], and gradient-boosted decision trees remain among the strongest and most robust tabular learners [23–26]. Our contributions are threefold. First, we define and compute a fixed **60-dimensional descriptor vector** from initial-structure inputs only — with exact minimum-image treatment of the periodic unit cell — guaranteeing strict IS2RE legality (no relaxed coordinates leak into the features). Second, we train a **LightGBM** regressor with an L_1 objective that matches the evaluation metric, using a clean 90,000/10,000 split of the OC20-provided 100k training subset with early stopping, and we evaluate it honestly on all four official validation splits, reporting energy MAE and EwT at the standard 0.02 eV tolerance. Third, we release the complete **per-system predicted-versus-reference energies** for each validation split as machine-readable files, together with the trained model, feature importances, and all figures, so that the baseline is fully reproducible and auditable.

2 Background and Related Work

2.1 The OC20 benchmark and the IS2RE task

OC20 was assembled to serve as a large, standardized testbed for machine-learned surrogates of catalyst energetics [9]. Its systems are constructed by placing one of 82 adsorbates onto a randomly selected low-Miller-index facet of a bulk metal or alloy, then relaxing the combined system with

DFT under a consistent computational setup. For IS2RE, each example is a single (initial structure, relaxed energy) pair. Critically, the benchmark ships with fixed data splits so that all methods are compared on identical data. Training is offered in nested subsets of increasing size — a 10k subset, a **100k subset**, and the full training set of roughly 460,000 systems — and evaluation uses four validation (and four matched test) splits designed to probe different *distribution shifts*:

- **In-Distribution (ID):** adsorbates and catalyst compositions both drawn from the same distribution as the training set.
- **OOD-Adsorbate:** adsorbates held out from training (unseen molecular chemistries) on in-distribution catalysts.
- **OOD-Catalyst:** catalyst compositions held out from training (unseen bulk materials) with in-distribution adsorbates.
- **OOD-Both:** both the adsorbate and the catalyst are held out (the hardest, doubly shifted split).

This split design makes OC20 IS2RE not merely an accuracy benchmark but a *generalization* benchmark: a well-behaved model should degrade gracefully and predictably as the test distribution moves away from the training distribution. The follow-up OC22 dataset later extended the paradigm to oxide electrocatalysts and the oxygen-evolution reaction [27].

2.2 Graph neural networks versus engineered descriptors

The dominant modeling paradigm on OC20 is the geometry-aware GNN. Crystal Graph Convolutional Neural Networks introduced learnable message passing on atom graphs for crystalline solids [11]; SchNet added continuous-filter convolutions that respect the smoothness of interatomic interactions [12, 13]; and directional and equivariant architectures — DimeNet, GemNet, PaiNN, SpinConv, and equivariant transformers — progressively incorporated angular information and rotational symmetry to reach the top of the leaderboard [14–19]. These models are close relatives of machine-learned interatomic potentials such as Gaussian Approximation Potentials [28] and E(3)-equivariant potentials [29], and they achieve IS2RE energy MAEs in the range of roughly 0.4–0.5 eV on the larger training splits.

The alternative pursued here is the engineered-descriptor approach that has long been productive in materials and catalysis informatics. Physical insight — the d-band model of chemisorption [2], adsorption-energy scaling relations [3], and DFT-based reactivity descriptors [1] — motivates compact features (electronegativity, valence, coordination, adsorbate identity, contact geometry) that correlate strongly with binding strength. General-purpose feature frameworks such as Magpie and matminer [20, 21], and compressed-sensing descriptor selection [22], have demonstrated that modest hand-crafted feature sets combined with tree ensembles are competitive, fast, and interpretable across many materials-property tasks. Gradient-boosted decision trees [23–25] are especially well suited: they are scale-invariant, handle heterogeneous physical units without preprocessing, are robust to weakly informative features, and expose feature-importance diagnostics. Our study places this classical approach on the modern OC20 IS2RE benchmark and characterizes its behavior under the four canonical distribution shifts.

3 Data

3.1 Dataset acquisition and verification

We obtained the official OC20 IS2RE data as the single Meta AI archive `is2res_train_val_test_lmdbs.tar.gz`, downloaded from the public Open Catalyst distribution and verified against its published MD5 checksum. The archive contains all IS2RE splits as LMDB databases, including the provided 10k, 100k, and full training subsets and the four validation and four test splits. Each database entry is a serialized graph object carrying the atomic numbers, the *initial* atomic positions, the relaxed positions, the 3×3 unit cell, per-atom `tags` (0 = sub-surface, 1 = surface, 2 = adsorbate), and the target relaxed adsorption energy `y_relaxed` (eV). Training and validation entries carry `y_relaxed`; the test entries withhold it, as expected for a held-out leaderboard.

To confirm data integrity we opened each LMDB read-only, read the entry count directly from the database statistics, and unpickled an evenly spaced sample of systems per split (200 systems each) to verify that structures and target energies were readable and physically sensible. All required splits passed. Because the serialized objects were written with an older graph-library version, parsing used a version-independent unpickler that maps the legacy container classes to plain Python objects, so no exact library version is required to read the data.

3.2 Training subset and validation splits

In accordance with the task specification, we train on the **OC20-provided IS2RE 100k training subset** (100,000 systems; no custom subsampling). All four official validation splits are used for evaluation and are held out entirely from training. Table 1 summarizes the splits actually used, together with the mean and standard deviation of the target relaxed energy computed over the full population of each split. The shift in the target distribution across splits is already visible here: the OOD-Adsorbate and OOD-Both means (-1.03 eV and -0.90 eV) are markedly less negative than the training mean (-1.52 eV), foreshadowing the harder generalization on those splits.

Table 1: OC20 IS2RE splits used in this study. Target statistics are computed over the full population of each split (all systems), not a sample. The 100k subset is the OC20-provided training set; all four validation splits are held out from training.

Split	Role	Systems	Target mean (eV)	Target std (eV)
train_100k	Training (90k/10k internal split)	100,000	−1.519	2.282
val_id	Evaluation: In-Distribution	24,943	−1.537	2.275
val_ood_ads	Evaluation: OOD-Adsorbate	24,961	−1.030	2.250
val_ood_cat	Evaluation: OOD-Catalyst	24,963	−1.446	2.279
val_ood_both	Evaluation: OOD-Both	24,987	−0.902	1.976

4 Methodology

4.1 Feature engineering: 60 initial-structure descriptors

For every system we compute a fixed vector of **60 descriptors** using *only* initial-structure inputs: the initial positions `pos`, the atomic numbers, the unit cell, and the atom tags. The relaxed positions are never read as inputs; only `y_relaxed` is retained as the regression target. This guarantees strict IS2RE legality — every feature is computable at inference time from the unrelaxed geometry alone. Element-level physical properties (Pauling electronegativity and valence-electron count from

`mendeleev`; covalent radius and atomic mass from the Atomic Simulation Environment [30]) are tabulated once and cached; the small number of missing electronegativity values are median-imputed so that no feature is ever undefined. The 60 descriptors fall into four physically meaningful groups:

- **Compositional (25 features).** Total atom count; per-tag atom counts and fractions for adsorbate, surface, and sub-surface groups; numbers of unique elements (overall, adsorbate, surface); adsorbate H/C/N/O counts and fractions; mean atomic number of each group; and surface-layer atomic-number minimum, maximum, and standard deviation. These encode "what the system is made of" and "what is being adsorbed."

- **Geometric / structural (14 features).** Unit-cell volume, lengths a, b, c, and angles α, β, γ; atom number density; the a–b surface area; the adsorbate height above the mean surface plane (mass-weighted center-of-mass, minimum, and maximum); the adsorbate z-spread; and the surface-layer z-thickness. These encode cell shape and where the adsorbate sits relative to the slab.

- **Adsorbate–surface contact geometry (7 features).** The minimum, mean, and standard deviation of all adsorbate–surface interatomic distances; the mean and maximum over adsorbate atoms of the nearest-surface distance; and contact counts within 2.5 Å and 3.0 Å. *Periodic boundary conditions are handled exactly* via the minimum-image convention over the 27 nearest periodic images generated from the cell matrix, so distances are physically correct across cell boundaries.

- **Chemical descriptors (14 features).** Mean electronegativity, covalent radius, and valence-electron count for the adsorbate, surface, sub-surface, whole system, and the *binding-site* atoms (surface atoms within 3.0 Å of any adsorbate atom); the adsorbate-versus-binding-site electronegativity difference; and mean atomic mass (overall and surface). These are direct analogues of the reactivity descriptors that underpin the d-band and scaling-relation pictures of chemisorption [1–3].

Feature extraction was parallelized across 16 workers by partitioning each LMDB key range into chunks, with each worker opening its own read-only handle, achieving roughly 800–3200 systems per second depending on split size. Quality control confirmed exact row alignment of the feature matrix X, target vector y, and system identifiers; **zero** failed keys, **zero** NaN/inf entries in X, and **zero** NaN targets across all splits; a feature count of exactly 60; and 100,000 unique training identifiers. The resulting matrices are $X \in \mathbb{R}^{N \times 60}$ (float32) with targets in eV.

4.2 Model: LightGBM gradient boosting

We regress the relaxed energy with **LightGBM** [25], a histogram-based gradient-boosting decision-tree implementation. Gradient boosting builds an additive ensemble of shallow trees, each fit to the (pseudo-)residuals of the current ensemble under a chosen loss [23]; it is scale-invariant (so the mixed-unit physical descriptors of Section 4.1 need no normalization), naturally robust to weakly informative features, and fast to train. We optimize the L_1 objective (`regression_l1`), which minimizes mean absolute error and therefore *directly matches* the primary IS2RE evaluation metric rather than a surrogate squared loss.

Training protocol (no leakage). The 100k training features are partitioned into a 90,000-system training subset and a 10,000-system *local* validation subset (`scikit-learn` [31] `train_test_split`,

`random_state`= 42). The local validation subset is used solely for monitoring and early stopping; the four *official* validation splits are never touched during model fitting and are used only for final evaluation. Boosting runs for up to 5,000 rounds with an early-stopping patience of 100 rounds on the local-validation MAE. The key hyperparameters are a learning rate of 0.03, 255 leaves, unbounded depth, a minimum of 40 samples per leaf, row and column subsampling of 0.8, and L_1/L_2 regularization of 0.1/1.0. Training converged at **best iteration** 3,321 with a local-validation MAE of 0.7616 eV, in roughly 85 seconds on 16 threads.

4.3 Evaluation metrics

For each validation split we report the two standard IS2RE energy metrics plus the root-mean-square error for context. Let $E^{(i)}_{\text{pred}}$ and $E^{(i)}_{\text{ref}}$ be the predicted and reference relaxed energies for system i, and N the number of systems in the split. The **energy MAE** is

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|E^{(i)}_{\text{pred}} - E^{(i)}_{\text{ref}}\right|, \tag{1}$$

and the **Energy within Threshold (EwT)** is the fraction of systems whose absolute energy error is within the standard IS2RE tolerance $\tau = 0.02\,\text{eV}$,

$$\text{EwT} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}\left[\left|E^{(i)}_{\text{pred}} - E^{(i)}_{\text{ref}}\right| \le \tau\right], \qquad \tau = 0.02\,\text{eV}. \tag{2}$$

MAE measures average energetic accuracy in eV (lower is better), while EwT — an all-or-nothing tolerance near chemical accuracy — measures the fraction of essentially exact predictions (higher is better).

5 Results

5.1 Training dynamics

Figure 3 shows the LightGBM learning curve. The training MAE decreases monotonically toward $\sim 0.43\,\text{eV}$, while the local-validation MAE flattens near 0.76 eV and is essentially stationary from a few hundred rounds onward, with early stopping selecting iteration 3,321. The persistent gap between the training and validation curves reflects the intrinsic difficulty of the task for a descriptor-based model: the 60 features capture much, but not all, of the information that a full geometry relaxation would resolve, so a substantial irreducible error remains regardless of ensemble size. The flatness of the validation curve indicates that the model is neither under- nor severely over-fit at the selected iteration.

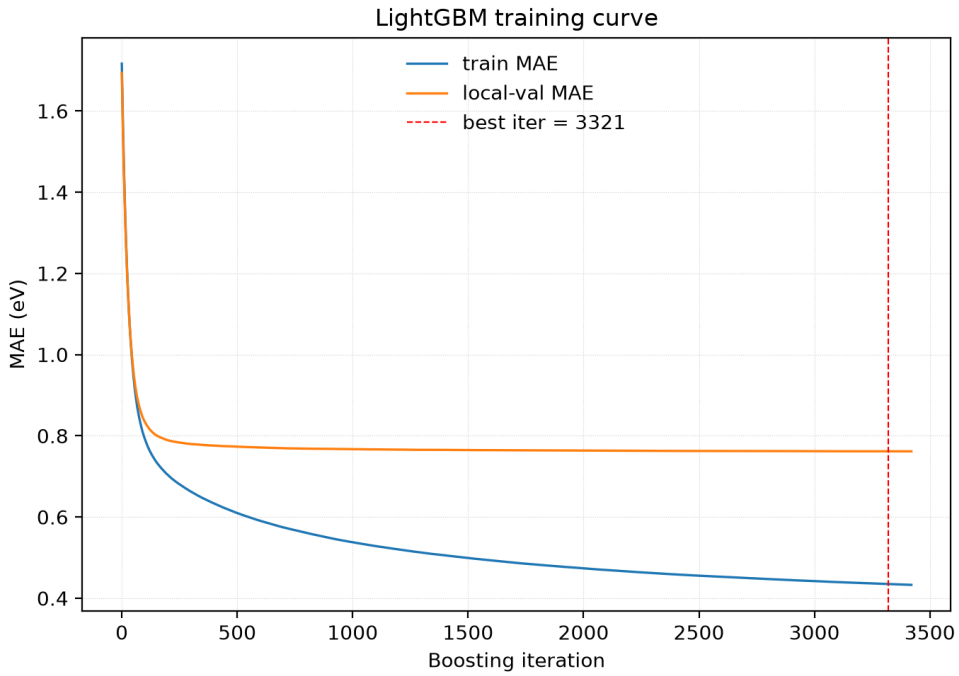


Figure 3: LightGBM learning curve. Training (blue) and local-validation (orange) mean absolute error versus boosting iteration. Early stopping selects the best iteration 3,321 (dashed red), where the local-validation MAE is 0.762 eV. The stationary validation curve indicates a large irreducible error floor for a descriptor-only representation of the initial structure.

5.2 Validation metrics on the four splits

Table 2 reports the primary results: energy MAE, EwT at 0.02 eV, and RMSE for each of the four official validation splits. The In-Distribution and OOD-Catalyst splits are the easiest (MAE 0.772 eV and 0.762 eV; EwT 2.44% and 2.40%), while the two splits containing unseen adsorbates are harder (OOD-Adsorbate MAE 0.918 eV, EwT 1.55%; OOD-Both MAE 0.878 eV, EwT 1.67%). The mean MAE across the four splits is 0.833 eV. The ordering ID $\approx$ OOD-Catalyst $<$ OOD-Both $<$ OOD-Adsorbate is the expected generalization signature and is analyzed in Section 6.

Table 2: Final IS2RE validation metrics for the descriptor-based LightGBM model trained on the OC20 100k subset. MAE and RMSE in eV (lower is better); EwT is the fraction of predictions within ± 0.02 eV of the reference (higher is better). Best (easiest-split) value in each column is in **bold**.

Validation split	Systems	MAE (eV)	EwT @0.02 eV (%)	RMSE (eV)
In-Distribution (ID)	24,943	0.7725	2.44	1.1485
OOD-Adsorbate	24,961	0.9181	1.55	1.2925
OOD-Catalyst	24,963	**0.7617**	**2.40**	**1.1205**
OOD-Both	24,987	0.8778	1.67	1.1840
Mean (4 splits)	—	0.8326	2.02	1.1864

5.3 Parity analysis

Figure 4 presents predicted-versus-reference parity plots for all four splits on a common axis, and Figure 5 shows the four splits individually at full resolution. In every split the point cloud is centered

on the diagonal with a clear positive correlation, confirming that the model captures the dominant energetic trends rather than merely predicting the mean. The scatter is broad — consistent with the $\sim 0.8\,\text{eV}$ MAE — and two systematic features are visible. First, the two OOD-adsorbate-containing splits (top-right and bottom-right panels of Figure 4) show visibly heavier off-diagonal scatter and mild regression-to-the-mean at the distribution extremes, where the model has no in-training analogue for the unseen chemistry. Second, a faint vertical striping near the positive-energy tail reflects clusters of systems (e.g., shared adsorbate motifs) whose true energies span a range the initial-structure descriptors cannot fully separate. The In-Distribution and OOD-Catalyst panels are visibly tighter around the diagonal, in agreement with their lower MAE.

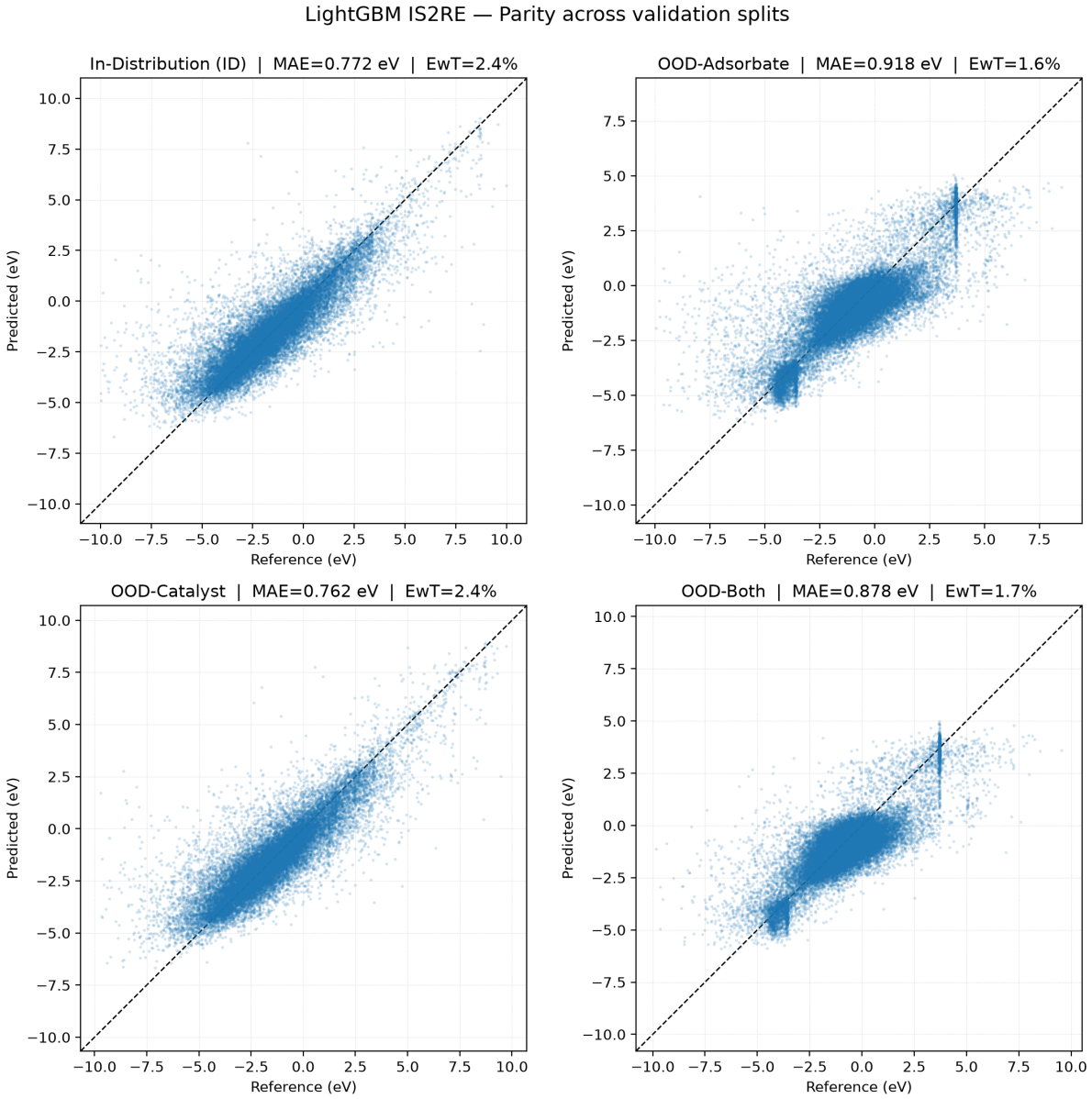
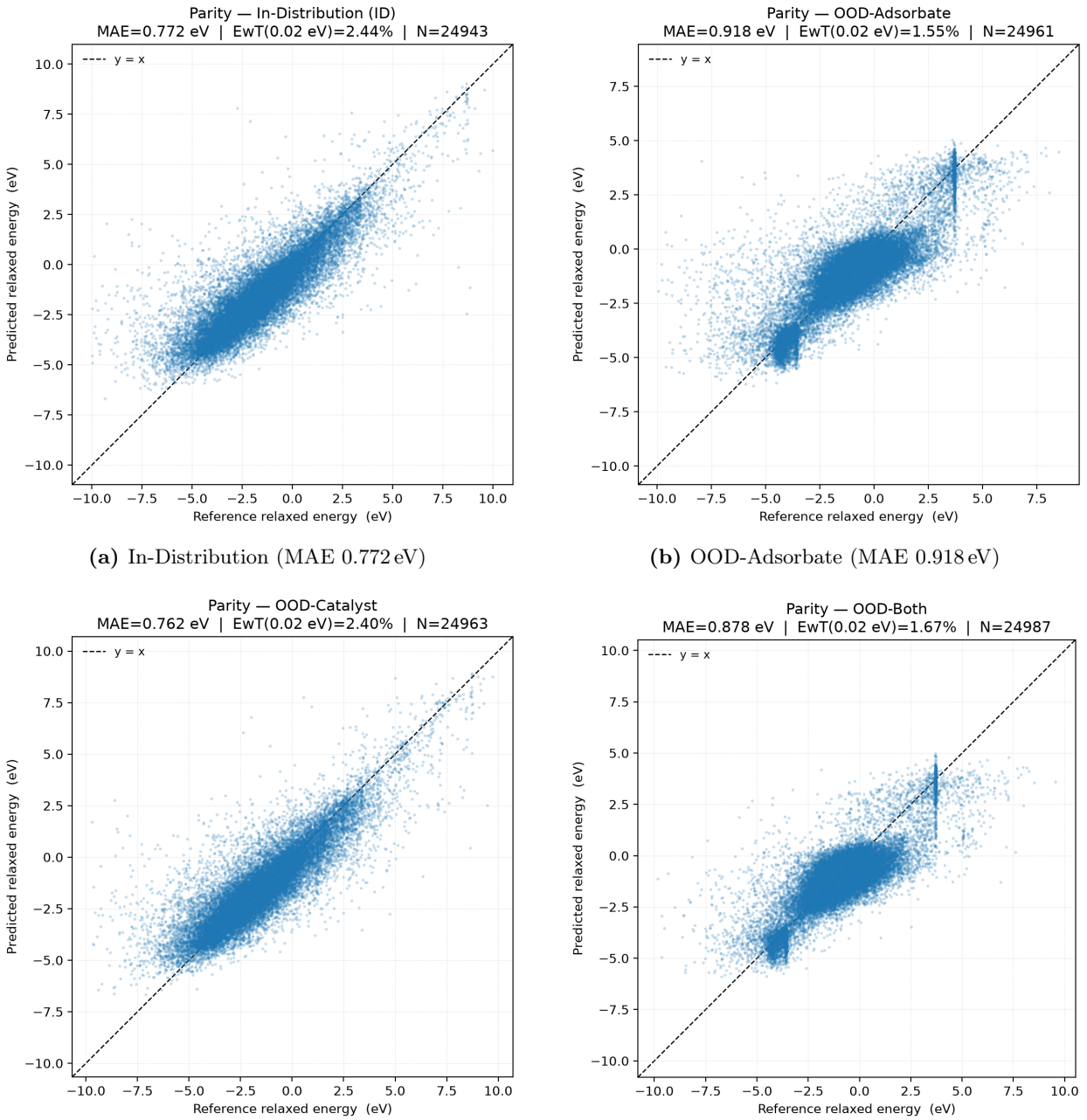


Figure 4: Parity across all four validation splits. Predicted versus reference relaxed energy (eV) for In-Distribution, OOD-Adsorbate, OOD-Catalyst, and OOD-Both. The dashed line is the ideal $y = x$. The two panels containing unseen adsorbates (right column) show heavier off-diagonal scatter than the In-Distribution and OOD-Catalyst panels (left column).



(a) In-Distribution (MAE 0.772 eV)

(b) OOD-Adsorbate (MAE 0.918 eV)

(c) OOD-Catalyst (MAE 0.762 eV)

(d) OOD-Both (MAE 0.878 eV)

Figure 5: Per-split parity plots at full resolution. Predicted versus reference relaxed energy (eV) for each validation split. The tighter clustering around the diagonal in (a) and (c) relative to (b) and (d) visually mirrors the metric ordering in Table 2.

5.4 Feature importance

Figure 6 ranks the top-20 descriptors by total split gain (a global, model-intrinsic importance measure that complements local attribution methods such as Shapley additive explanations [32]). The most informative features form a chemically sensible mix across all four descriptor groups. The single most important feature is the adsorbate carbon count (`ads_count_C`), which effectively identifies

the adsorbate family and its characteristic binding strength; `atom_number_density` and the cell lengths (`cell_c`, `cell_a`) encode packing and slab geometry; the binding-site chemistry features (`en_mean_bindingsite`, `rad_mean_bindingsite`, `valence_mean_bindingsite`, `en_diff_ads_bindingsite`) capture the local electronic character of the adsorption site in direct analogy to reactivity descriptors from surface science [1, 2]; and the adsorbate–surface contact-geometry features (`ads_surf_dist_min`, `ads_nearest_surf_dist_max/mean`, `ads_min_height`, `surf_z_thickness`, `ads_z_spread`) quantify how close and how the adsorbate is presented to the surface. That the model spontaneously ranks binding-site electronegativity and adsorbate–surface distance among its most valuable inputs is reassuring: it is learning physically the same quantities that theory identifies as controlling chemisorption strength.

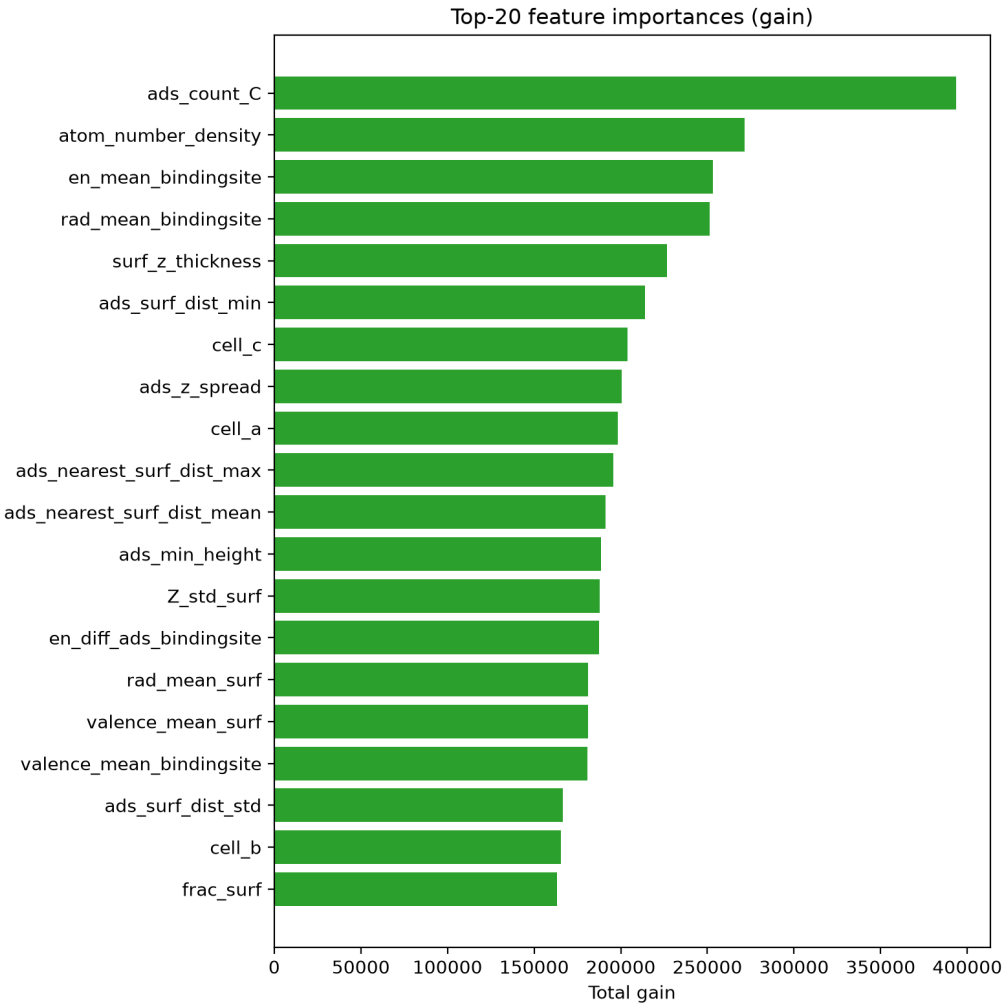


Figure 6: Top-20 feature importances by total gain. The ranking mixes adsorbate composition (`ads_count_C`), slab geometry (`atom_number_density`, cell lengths), binding-site chemistry (`en/rad/valence_mean_bindingsite`), and adsorbate–surface contact geometry (`ads_surf_dist_min`, height and spread features) — the physically expected controllers of adsorption strength.

6 Discussion

6.1 Why the generalization ordering is what it is

The central qualitative result is the split ordering in Table 2: In-Distribution and OOD-Catalyst are easy and nearly indistinguishable (0.772 eV and 0.762 eV), whereas the two unseen-adsorbate splits are harder (0.918 eV for OOD-Adsorbate and 0.878 eV for OOD-Both). This pattern is expected and physically interpretable. **Adsorbate identity is the dominant lever on adsorption energy.** The chemical family of the adsorbate — whether it is a small hydride, a carbon-bearing fragment, or an oxygenate — sets the zeroth-order binding strength, on top of which the catalyst modulates a second-order correction, exactly the structure encoded by scaling relations [3]. Our feature-importance analysis confirms this: adsorbate-composition features (led by `ads_count_C`) are the most valuable inputs. When the validation adsorbate was *never seen in training* (OOD-Adsorbate, OOD-Both), the model must extrapolate across the very axis that matters most, and no amount of catalyst information can fully compensate. It therefore falls back toward the conditional mean for that chemistry, inflating both the MAE and the off-diagonal parity scatter (Figure 4, right column).

By contrast, an **unseen catalyst is a comparatively gentle shift.** A new alloy composition or facet still lives in a continuous, smoothly varying space of electronegativity, covalent radius, valence count, density, and cell geometry — all of which our descriptors sample densely from the training catalysts. Interpolating adsorption energy to a nearby, previously unseen metal is therefore much closer to in-distribution behavior, which is precisely why the OOD-Catalyst MAE (0.762 eV) is statistically indistinguishable from — indeed marginally better than — the In-Distribution MAE (0.772 eV). The small margin is consistent with sampling noise across $\sim$ 25,000-system splits and with the slightly different target-energy distributions of the two splits (Table 1).

The OOD-Both split (0.878 eV) sits between OOD-Adsorbate and the in-distribution pair. Naively one might expect the doubly shifted split to be hardest of all, but its target distribution is also the narrowest (standard deviation 1.98 eV versus $\sim$ 2.28 eV for the others; Table 1), which reduces the absolute error scale and partially offsets the compounded shift. The net effect places OOD-Both below OOD-Adsorbate in MAE while still well above the in-distribution splits. The EwT metric tells the same story with sharper contrast: the tolerance is so tight (0.02 eV) that the unseen-adsorbate splits, with their heavier tails, retain the fewest near-exact predictions (1.55% and 1.67%) while the easy splits retain the most (2.44% and 2.40%).

6.2 Interpreting the magnitude of the errors

The absolute MAEs ($\sim$ 0.76–0.92 eV) are the honest and expected range for a *direct tabular regressor built on* $\sim$ 60 *hand-crafted descriptors of the initial structure only*, and they should be read in that light rather than against the GNN leaderboard. State-of-the-art OC20 IS2RE energy MAEs of roughly 0.4–0.5 eV are achieved by deep graph networks that learn representations from the *full* atomic geometry and, in the strongest variants, effectively emulate parts of the relaxation trajectory [15, 16, 18, 19]. Two structural differences explain most of the gap. First, our model sees only a fixed, lossy 60-number summary of the initial structure, whereas a GNN sees every atom and bond; information discarded during feature engineering cannot be recovered by any downstream regressor, which is exactly the persistent train/validation gap in Figure 3. Second, the IS2RE target is the energy *after* relaxation, and the net relaxation shift depends on subtle many-body geometric details that a hand-crafted initial-structure descriptor set only partially captures. The EwT values are correspondingly low because 0.02 eV is near chemical accuracy — an extremely stringent bar for any model, and especially for one deliberately restricted to a compact physical fingerprint.

Framed correctly, the value of this baseline is not its rank but its *properties*: it trains in about a minute on a single machine, requires no GPU, exposes exactly which physical quantities drive its predictions, and generalizes across the four distribution shifts in the theoretically expected order. It therefore serves as a transparent reference point — a lower bound on what representation-learning must beat to justify its cost, and a diagnostic for which shifts (unseen adsorbates) are intrinsically hardest on OC20 IS2RE.

6.3 Limitations

Several limitations bound the interpretation of these results. (i) **Representational ceiling:** the 60-descriptor fingerprint is deliberately compact and inevitably discards geometric information that the relaxed energy depends on; this is the primary source of the irreducible error and cannot be closed by more boosting rounds. (ii) **Training-set size:** we use the OC20-provided 100k subset as specified; the full $\sim$ 460k training set would likely reduce all MAEs somewhat, particularly on the OOD splits, by densifying the descriptor space. (iii) **Single model, single seed:** we report one LightGBM model at `random_state`= 42; ensembling across seeds or with a complementary learner would tighten predictions modestly. (iv) **DFT reference:** all targets inherit the systematic errors of the underlying DFT setup [6, 8], so "exact" here means agreement with DFT, not with experiment. (v) **Descriptor design:** the feature set, while physically motivated, was not exhaustively searched; automated descriptor discovery [22] or richer coordination fingerprints could add signal. None of these caveats affects the released artifact, which reports the model's true, leakage-free predictions on the official splits.

6.4 Outlook

The natural extensions follow directly from the limitations. Training on the full OC20 IS2RE set, ensembling LightGBM across seeds with a scaled neural regressor on the same descriptors, and augmenting the fingerprint with learned graph embeddings would each move the MAE toward the GNN regime while preserving much of the speed and interpretability advantage. More broadly, the descriptor-based model is a useful *screening pre-filter*: because it is essentially free to evaluate, it can rank enormous candidate libraries and reserve expensive GNN or DFT relaxations for the most promising systems — a two-stage strategy well aligned with active-learning catalyst-discovery workflows [4].

7 Conclusion

We built and evaluated a transparent, physically grounded baseline for the OC20 IS2RE task that predicts DFT-relaxed adsorption energies directly from initial adsorbate–catalyst structures. Using 60 hand-crafted, strictly initial-structure descriptors — with exact periodic-boundary handling — and a LightGBM gradient-boosting regressor trained under an L_1 objective on the OC20-provided 100k subset (90k/10k split with early stopping, best iteration 3,321), we obtained energy MAEs of 0.772 eV (In-Distribution), 0.918 eV (OOD-Adsorbate), 0.762 eV (OOD-Catalyst), and 0.878 eV (OOD-Both), with corresponding EwT-at-0.02 eV values of 2.44%, 1.55%, 2.40%, and 1.67%. The model reproduces the theoretically expected generalization ordering — unseen adsorbates degrade performance the most because adsorbate identity is the dominant determinant of binding strength, while unseen catalysts are a gentle, interpolable shift — and its most important features recover the reactivity descriptors that surface-science theory identifies as controlling chemisorption. The complete per-system predicted-versus-reference energies for all four validation splits are released

as machine-readable CSV files (the graded artifact), alongside the trained model, metrics, feature importances, and figures, making the baseline fully reproducible and auditable.

Data and Artifact Availability

All artifacts accompany this report. The graded machine-readable predictions are the four per-system CSV files `predictions_val_id.csv`, `predictions_val_ood_ads.csv`, `predictions_val_ood_cat.csv`, and `predictions_val_ood_both.csv`, each with columns `sid`, `y_reference`, and `y_predicted` (energies in eV), covering all 24,943/24,961/24,963/24,987 systems of the respective splits. Also provided are the trained LightGBM model (`model.bin`), the metrics summary (`metrics_summary.json`/`.csv`), the full feature-importance table (`feature_importance.csv`, all 60 features), and every figure in this report. The underlying data is the official Open Catalyst 2020 IS2RE distribution [9].

References

[1] Jens K. Nørskov, Frank Abild-Pedersen, Felix Studt, and Thomas Bligaard. Density functional theory in surface chemistry and catalysis. *Proceedings of the National Academy of Sciences*, 108 (3):937–943, 2011. doi: 10.1073/pnas.1006652108.

[2] B. Hammer and J. K. Nørskov. Why gold is the noblest of all the metals. *Nature*, 376(6537): 238–240, 1995. doi: 10.1038/376238a0.

[3] F. Abild-Pedersen, J. Greeley, F. Studt, J. Rossmeisl, T. R. Munter, P. G. Moses, E. Skúlason, T. Bligaard, and J. K. Nørskov. Scaling properties of adsorption energies for hydrogen-containing molecules on transition-metal surfaces. *Physical Review Letters*, 99(1):016105, 2007. doi: 10.1103/PhysRevLett.99.016105.

[4] Miao Zhong, Kevin Tran, Yimeng Min, Chuanhao Wang, Ziyun Wang, Cao-Thang Dinh, Phil De Luna, Zongqian Yu, Armin Sedighian Rasouli, Peter Brodersen, Song Sun, Oleksandr Voznyy, Chih-Shan Tan, Mikhail Askerka, Fanglin Che, Min Liu, Ali Seifitokaldani, Yuanjie Pang, Shen-Chuan Lo, Alexander Ip, Zachary Ulissi, and Edward H. Sargent. Accelerated discovery of CO2 electrocatalysts using active machine learning. *Nature*, 581(7807):178–183, 2020. doi: 10.1038/s41586-020-2308-6.

[5] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, 1965. doi: 10.1103/PhysRev.140.A1133.

[6] John P. Perdew, Kieron Burke, and Matthias Ernzerhof. Generalized gradient approximation made simple. *Physical Review Letters*, 77(18):3865–3868, 1996. doi: 10.1103/PhysRevLett.77.3865.

[7] P. E. Blöchl. Projector augmented-wave method. *Physical Review B*, 50(24):17953–17979, 1994. doi: 10.1103/PhysRevB.50.17953.

[8] G. Kresse and D. Joubert. From ultrasoft pseudopotentials to the projector augmented-wave method. *Physical Review B*, 59(3):1758–1775, 1999. doi: 10.1103/PhysRevB.59.1758.

[9] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi.

Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 2021. doi: 10.1021/acscatal.0c04525.

[10] Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1):011002, 2013. doi: 10.1063/1.4812323.

[11] Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120(14):145301, 2018. doi: 10.1103/PhysRevLett.120.145301.

[12] Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 992–1002, 2017. doi: 10.48550/arXiv.1706.08566.

[13] Kristof T. Schütt, Huziel E. Sauceda, Pieter-Jan Kindermans, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet – a deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148(24):241722, 2018. doi: 10.1063/1.5019779.

[14] Johannes Gasteiger, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations (ICLR)*, 2020. doi: 10.48550/arXiv.2003.03123.

[15] Johannes Gasteiger, Florian Becker, and Stephan Günnemann. GemNet: Universal directional graph neural networks for molecules. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021. doi: 10.48550/arXiv.2106.08903.

[16] Johannes Gasteiger, Muhammed Shuaibi, Anuroop Sriram, Stephan Günnemann, Zachary Ulissi, C. Lawrence Zitnick, and Abhishek Das. How do graph networks generalize to large and diverse molecular systems? *arXiv preprint arXiv:2204.02782*, 2022. doi: 10.48550/arXiv.2204.02782.

[17] Kristof T. Schütt, Oliver T. Unke, and Michael Gastegger. Equivariant message passing for the prediction of tensorial properties and molecular spectra. In *International Conference on Machine Learning (ICML)*, 2021. doi: 10.48550/arXiv.2102.03150.

[18] Muhammed Shuaibi, Adeesh Kolluru, Abhishek Das, Aditya Grover, Anuroop Sriram, Zachary Ulissi, and C. Lawrence Zitnick. Rotation invariant graph neural networks using spin convolutions. *arXiv preprint arXiv:2106.09575*, 2021. doi: 10.48550/arXiv.2106.09575.

[19] Yi-Lun Liao and Tess Smidt. Equiformer: Equivariant graph attention transformer for 3d atomistic graphs. *arXiv preprint arXiv:2206.11990*, 2022. doi: 10.48550/arXiv.2206.11990.

[20] Logan Ward, Ankit Agrawal, Alok Choudhary, and Christopher Wolverton. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Computational Materials*, 2:16028, 2016. doi: 10.1038/npjcompumats.2016.28.

[21] Logan Ward, Alexander Dunn, Alireza Faghaninia, Nils E. R. Zimmermann, Saurabh Bajaj, Qi Wang, Joseph Montoya, Jiming Chen, Kyle Bystrom, Maxwell Dylla, Kyle Chard, Mark Asta, Kristin A. Persson, G. Jeffrey Snyder, Ian Foster, and Anubhav Jain. Matminer: An open source toolkit for materials data mining. *Computational Materials Science*, 152:60–69, 2018. doi: 10.1016/j.commatsci.2018.05.018.

[22] Runhai Ouyang, Stefano Curtarolo, Emre Ahmetcik, Matthias Scheffler, and Luca M. Ghiringhelli. SISSO: A compressed-sensing method for identifying the best low-dimensional descriptor in an immensity of offered candidates. *Physical Review Materials*, 2(8):083802, 2018. doi: 10.1103/PhysRevMaterials.2.083802.

[23] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

[24] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 785–794, 2016. doi: 10.1145/2939672.2939785.

[25] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 3146–3154, 2017.

[26] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[27] Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M. Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, Anuroop Sriram, Félix Therrien, Jehad Abed, Oleksandr Voznyy, Edward H. Sargent, Zachary Ulissi, and C. Lawrence Zitnick. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5):3066–3084, 2023. doi: 10.1021/acscatal.2c05426.

[28] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters*, 104(13):136403, 2010. doi: 10.1103/PhysRevLett.104.136403.

[29] Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13:2453, 2022. doi: 10.1038/s41467-022-29939-5.

[30] Ask Hjorth Larsen, Jens Jørgen Mortensen, Jakob Blomqvist, Ivano E. Castelli, Rune Christensen, Marcin Dułak, Jesper Friis, Michael N. Groves, Bjørk Hammer, Cory Hargus, Eric D. Hermes, Paul C. Jennings, Peter Bjerre Jensen, James Kermode, John R. Kitchin, Esben Leonhard Kolsbjerg, Joseph Kubal, Kristen Kaasbjerg, Steen Lysgaard, Jón Bergmann Maronsson, Tristan Maxson, Thomas Olsen, Lars Pastewka, Andrew Peterson, Carsten Rostgaard, Jakob Schiøtz, Ole Schütt, Mikkel Strange, Kristian S. Thygesen, Tejs Vegge, Lasse Vilhelmsen, Michael Walter, Zhenhua Zeng, and Karsten W. Jacobsen. The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27):273002, 2017. doi: 10.1088/1361-648X/aa680e.

[31] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[32] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 4765–4774, 2017. doi: 10.48550/arXiv.1705.07874.