RISK-CONTROLLED DFT SELECTION FOR MACHINE-LEARNED INTERATOMIC POTENTIALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Density functional theory (DFT) gives accurate energies and forces but is expensive. Pretrained machine-learned interatomic potentials (MLIPs) are widely used for their cheap costs, but their errors vary widely across molecules. So how do we choose which configurations to trust with MLIP and which to send to DFT? We cast this as risk-controlled selection and solve it with a conformal risk control (CRC) framework. A finite CRC rule turns any error score fixed before calibration into a routing decision, whose expected capped per-molecule error left to the MLIP is at most a user-chosen fraction α of its fitting-set mean. Because every score receives the same guarantee, scores can be compared by the DFT cost or call. With this comparison across eight dataset–MLIP pairs, we find that committee disagreement often overlooks errors that the models share. We also show that an analytical-chemistry-informed lightweight error predictor ranks errors more accurately than existing error-score baselines. At $\alpha = 0.3$ it sends 7% and 14% fewer than two committee-based baselines while trailing by 3.4 percentage points from oracle. The advantage is emphasized when the user's DFT reference differs from the MLIPs' training reference and is less apparent when the references match. We also show that replacing only some MLIP energies with DFT can break error cancellation in energy differences, so guarantees on relative energies must be calibrated directly.

1 INTRODUCTION

Density functional theory (DFT) provides accurate energies and atomic forces from electronic structure and is a central computational tool across chemistry and materials science (Kohn & Sham, 1965). However, conventional DFT scales approximately as $O(N^3)$ with system size, where N denotes the number of atoms, which limits the number of atomic configurations that can be evaluated in large-scale applications such as molecular dynamics and high-throughput screening (Bowler & Miyazaki, 2012; Curtarolo et al., 2013). Machine-learned interatomic potentials (MLIPs) address this bottleneck by approximating DFT energies and forces at a small fraction of the computational cost, typically with linear scaling in N for local architectures (Behler & Parrinello, 2007; Unke et al., 2021; Batatia et al., 2022). Larger training datasets and increased model capacity have substantially improved the accuracy and generality of MLIPs, making pretrained models practical surrogates for DFT across broad regions of chemical space (Batatia et al., 2025; Wood et al., 2025). However, no pretrained MLIP is uniformly accurate, and prediction errors can vary substantially across molecules, geometries, and chemical environments (Fu et al., 2023; Deng et al., 2025; Perez et al., 2025).

This creates a basic decision problem whenever an MLIP should be used to accelerate a DFT workflow. The main question then becomes: Given a large pool of molecular configurations, which configurations can safely be evaluated with the MLIP and which should be sent to DFT? Evaluating every configuration with DFT increases the computational cost, while accepting every MLIP prediction results in configurations with large energy or force errors. A natural strategy is to estimate the error for each configuration and send only the unreliable configurations to DFT (Jinnouchi et al., 2019; Liu et al., 2026). The most common estimate is the disagreement within a committee of MLIPs (Smith et al., 2018; Schran et al., 2020). However, disagreement only flags errors on which the models differ, and errors that all models share remain invisible. The challenge is that the true error is unavailable without performing the expensive DFT calculation (Perez et al., 2025). Moreover, an error predictor that ranks configurations well does not by itself provide a finite-sample guarantee

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

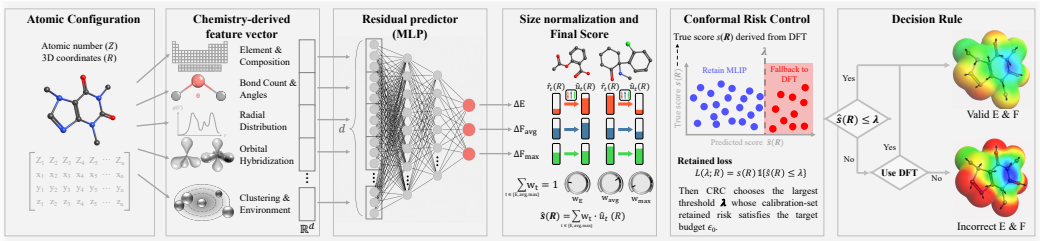

Figure 1: Overview of the MLIP–DFT selection framework. A molecular configuration (caffeine in this example) is converted into a chemistry-derived structural vector and passes through a lightweight predictor to estimate energy and force residuals. These estimates are normalized and combined into a predicted error score $\hat{s}(R)$. Conformal risk control then calibrates a threshold that retains reliable MLIP predictions and sends higher-risk configurations to DFT. The score and the capped loss are fixed before calibration.

on the error remaining among configurations that are not evaluated with DFT (Angelopoulos et al., 2024). Predicting error is not the same as knowing where to stop trusting the MLIP.

In this work, we treat MLIP–DFT routing as risk-controlled selection for a frozen pretrained MLIP and solve it using conformal risk control (CRC) (Bates et al., 2021; Angelopoulos et al., 2024) (Figure 1). The method separates ranking from deciding. For ranking, a lightweight predictor maps an analytical chemistry consensus-rooted structural representation to the expected MLIP–DFT error in three channels: energy, average force, and maximum force. The channels are combined into a single predicted error score. For deciding, CRC uses a labeled calibration set to turn the score into a threshold, and every configuration scoring above this threshold is sent to DFT. Note the exchangeable unit of calibration is a *molecule identity* (the set of all sampled geometries of one molecule). Routing decisions, however, are made for individual *configurations*. Calibration controls validity. Under exchangeability of molecule identities and for any error score fixed before calibration, the expected capped error left in the MLIP-evaluated configurations of a test identity is at most a user-chosen fraction α of the mean per-identity error on the fitting set. The quality of the ranking controls only efficiency, that is, how many DFT calls are needed to meet this budget. This separation is the organizing principle of the paper: the error predictor ranks by uncertainty, and CRC decides where to stop. Because every score fixed before calibration receives the same guarantee, the DFT spent by the score becomes a common metric for comparing reliability signals. A better ranking can thus lower DFT cost without changing the risk-control layer, if fixed before calibration.

Our **contributions** are as follows: (i) **Risk-controlled routing.** We cast MLIP–DFT routing as conformal risk control over molecular identities. For any score fixed before calibration, the expected capped error left to the MLIP stays within a user-set budget, and ranking quality only sets the DFT cost (Section 2.4). (ii) **A model-agnostic error ranker.** A lightweight predictor on a structure-only, chemistry-informed descriptor ranks MLIP errors about as well as the target MLIP's own latent representation (Sections 2.2–2.3). (iii) **What reliability signals miss.** Under the same guarantee, committee disagreement misses errors shared across MLIPs. Structure-based ranking needs less DFT when the user's DFT reference differs from the MLIPs' training reference, and the difference closes when they match. Partial DFT replacement can also break error cancellation in energy differences (Section 3.2, Appendix J.1).

Related work. The closest precedents are Liu et al. (2026), who send structures to DFT using an empirically chosen cutoff on ensemble uncertainty, and Conformal Arbitrage (Overman & Bayati, 2025), which also calibrates a deferral threshold with CRC; we differ in using a separately trained error predictor and in controlling an identity-level loss over per-configuration decisions. Appendix A gives a full comparison.

2 METHOD

2.1 SETUP AND NOTATION

Let R denote a configuration with $N = N(R)$ atoms, atomic numbers $Z_j(R)$ and positions $\mathbf{x}_j(R) \in \mathbb{R}^3$, $j = 1, \ldots, N$. DFT assigns R an energy $e^*(R) \in \mathbb{R}$ and per-atom forces $\mathbf{f}_j^*(R) \in \mathbb{R}^3$; a

pretrained MLIP model outputs predictions $\hat{e}(R)$ and $\hat{\mathbf{f}}_j(R)$ cheaply. MLIP leaves residuals behind,

$$\delta(R) = e^*(R) - \hat{e}(R) \in \mathbb{R}, \qquad \mathbf{g}_j(R) = \mathbf{f}_j^*(R) - \hat{\mathbf{f}}_j(R) \in \mathbb{R}^3, \quad j = 1, \ldots, N. \tag{1}$$

A frozen public MLIP and a reference DFT dataset do not share an isolated-atom energy convention, so we work with the aligned energy residual $\Delta E(R) = \delta(R) - \phi(R)^\top \hat{a}$, where $\phi(R)$ holds element counts and an intercept and $\hat{a}$ is a composition-linear fit on $\mathcal{D}_{\text{fit}}$ that is frozen before calibration (Appendix D). Alignment leaves force residuals and within-composition energy differences unchanged. Here, DFT means the reference the user trusts, which may use a different functional or basis set than MLIP's training data, and the residual therefore includes this difference.

The screening problem. Running DFT on a configuration drives its residual to zero, but costs far more than an MLIP forward pass. Given a frozen MLIP and an unlabeled pool of configurations, our goal is to send only a small uncertain subset to DFT and to accept MLIP predictions for the rest. However, a problem arises: labeling a configuration's residual requires the very DFT calculation we are trying to avoid. We therefore need a label-free proxy for the residual, together with a principled rule for how large that proxy must be before a DFT calculation is worth its cost.

Data splits. We use three disjoint sets. $\mathcal{D}_{\text{fit}}$ and $\mathcal{D}_{\text{calib}}$ are DFT-labeled molecule identities that the MLIP was not refit on, used to fit the MLIP error predictor (Section 2.3) and to calibrate the selection cutoff (Section 2.4), respectively. $\mathcal{D}_{\text{test}}$ is the unlabeled test pool where the selection rule is applied. All configurations of a molecule identity stay in one split (Section 2.4).

Our method has two steps. First, we use the DFT-labeled set $\mathcal{D}_{\text{fit}}$ to define a single error score that combines the MLIP energy error and force residuals. Then we train a lightweight predictor $\hat{s}$ to estimate that error score from label-free structural information alone, and every configuration in $\mathcal{D}_{\text{test}}$ receives an estimated error score. Second, conformal risk control uses $\mathcal{D}_{\text{calib}}$ to convert those scores into a cutoff, and every pool configuration scoring above the cutoff is sent to DFT (Section 2.4). Sections 2.2 and 2.3 describe the two halves of the first step, and Section 2.4 explains the cutoff and its guarantee.

2.2 Structural vector

We define a structural vector constructed using only atomic numbers Z_j and Cartesian coordinates $\mathbf{x}_j$. No other features, including DFT quantities, learned embeddings, or MLIP features, enter its construction. The feature families are motivated by structural concepts from physical and analytical chemistry (Musil et al., 2021) and by established fixed-form descriptors such as atom-centered symmetry functions, the Coulomb matrix, and SOAP (Behler, 2011; Rupp et al., 2012; Bartók et al., 2013). They encode composition, local coordination, pair-distance distortion, short-range crowding, angular distortion, radial structure, and elemental composition of neighboring coordination shells. These are geometric variables that chemists routinely use to reason about molecular structure from nuclear magnetic resonance, Fourier-transform infrared spectroscopy, and Raman spectroscopy (Geue et al., 2022; Stavitski & Weckhuysen, 2010).

The descriptor concatenates nine fixed feature families into 350 dimensions (Table 2 in Appendix B). Coarse, discrete families count elements, element-resolved coordination states, and compressed, nominal, or stretched element-pair contacts. Continuous families then resolve what the counts hide: atomic-number-weighted short contacts, the few worst local environments, a smooth radial distribution, unweighted inverse-distance proximity at several powers, soft crowding across four distance scales, and the elemental identity of each atom's neighborhood. Exact definitions are given there.

2.3 A joint energy–force error predictor

Using the structural vector of Section 2.2 as the input, we train a lightweight error predictor that estimates how accurate the MLIP energy and force prediction is. For a configuration R, the predictor estimates three targets: energy residual, average force residual, and maximum force residual,

$$u_E(R) = |\Delta E(R)|, \qquad u_{\text{avg}}(R) = \frac{1}{N}\sum_{j=1}^{N} \|\mathbf{g}_j(R)\|_2, \qquad u_{\max}(R) = \max_j \|\mathbf{g}_j(R)\|_2, \tag{2}$$

where $\Delta E(R)$ and $\mathbf{g}_j(R)$ are the aligned energy and per-atom force residuals of (1). The three targets capture different failure modes, each of which matters in practical molecular dynamics (MD) (Fu et al., 2023). u_E detects a wrong total energy, u_{avg} detects forces that are broadly inaccurate and mislead the trajectory, and $u_{\max}$ detects a single badly mispredicted atom that crashes the simulation. Because the three residual targets scale differently with atom count N, each is divided by a size-dependent scale $\hat{c}_t N^{\hat{\beta}_t}$ fitted on $\mathcal{D}_{\mathrm{fit}}$ and then frozen, detailed in Appendix D; we write u_t for the normalized targets and $\hat{u}_t$ for their predictions from now on. The normalized residuals are combined into a single error score:

$$\hat{s}(R) = \sum_t w_t \hat{u}_t(R), \qquad s(R) = \sum_t w_t u_t(R), \qquad w_t \geq 0, \qquad \sum_t w_t = 1. \tag{3}$$

Here, $\hat{s}(R)$ is available from structure alone, whereas $s(R)$ requires the DFT reference label. The weights are fixed before calibration and specify the relative importance of energy, average-force, and maximum-force errors for the intended application—e.g. a user who cares more about force accuracy than energy accuracy can set w accordingly. All experiments in this paper use equal weights $w_t = 1/3$.

Fitting the predictor. Routing depends on the predictor only through the order it induces, so we ask it to rank well, not merely to regress well. The prediction $\hat{u}_t$ is the average of three small multilayer perceptrons (seeds 0, 1, 2; two hidden layers of 256 units) trained independently on $\mathcal{D}_{\mathrm{fit}}$. Each network b minimizes an ℓ_1 residual loss on the three normalized channels plus a pairwise ranking loss on its own composite score $q_b(R) = \sum_t w_t \hat{u}_{b,t}(R)$. The ranking targets are the truncated error scores $\ell_i = \ell(R_i)$ of Eq. (5) in Section 2.4, the same quantity that CRC controls. For a mini-batch $\mathcal{B}$ with a uniformly permuted pairing π, let $\mathcal{P} = \{i \in \mathcal{B} : \ell_i \neq \ell_{\pi(i)}\}$,

$$\mathcal{L}_b = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_t w_t \big|\hat{u}_{b,t}(R_i) - u_t(R_i)\big| + \frac{\lambda_{\mathrm{rank}}}{|\mathcal{P}|} \sum_{i \in \mathcal{P}} \log\Big[1 + e^{-(q_b(R_i) - q_b(R_{\pi(i)}))\ \mathrm{sign}(\ell_i - \ell_{\pi(i)})}\Big], \tag{4}$$

with $\lambda_{\mathrm{rank}} = 1$ and ℓ the truncated loss of Section 2.4. The ranking loss term pushes configurations with larger true loss toward larger scores; the regression term keeps the three channels interpretable. Architecture, seeds, and the value of λ_{rank} were chosen on a validation split internal to $\mathcal{D}_{\mathrm{fit}}$ and frozen before calibration (Appendix I).

2.4 Molecule selection by conformal risk control

The error predictor $\hat{s}$ can rank configurations but carries no guarantee about the true error of the unselected—even a perfect ranking would not say where to stop, since that is a decision about tolerable MLIP error against DFT cost, not a property of the predictor. CRC sidesteps this: the user instead chooses α, the fraction of the mean per-identity MLIP error they are willing to tolerate, and calibration returns the cutoff $\hat{\lambda}$ on $\hat{s}$ that achieves it.

This exposes an explicit tradeoff between *retained MLIP error* and *DFT calls*—tolerating more error means fewer calls. Sweeping α traces this cost–accuracy curve (Fig. 2, middle row), and the practitioner picks an operating point on it rather than guessing one in advance.

We follow the conformal risk control (CRC) framework of Angelopoulos et al. (2024), with one adaptation: the exchangeable unit is a molecule identity, not a configuration (Dunn et al., 2023). We write g for an identity and $\mathcal{R}_g$ for its set of configurations, and assign whole identities to splits $\mathcal{D}_{\mathrm{fit}}, \mathcal{D}_{\mathrm{calib}}, \mathcal{D}_{\mathrm{test}}$ uniformly at random. When a dataset does not record the parent identity of a configuration, that configuration forms its own group; see Appendix D.

Assumption 1 (Exchangeability of molecule identities)**.** The calibration identities $\mathcal{D}_{\mathrm{calib}} = \{g_i\}_{i=1}^{n}$ and the test identities $\mathcal{D}_{\mathrm{test}} = \{g_i\}_{i=n+1}^{n+m}$, each carrying its finite configuration set $\mathcal{R}_{g_i}$ with finite scores, are exchangeable.

Selection, on the other hand, still happens one configuration at a time—a molecule identity may have some of its configurations sent to DFT and the rest left to the MLIP. Let $s(R)$ be the true error score of Section 2.3 and $\hat{s}(R)$ its estimate. CRC requires a bounded loss (Bates et al., 2021; Angelopoulos et al., 2024), but s is commonly heavy-tailed, so we fix a truncation level $B_{\mathrm{cfg}} > 0$ and set

$$\ell(R) = \min\{s(R),\, B_{\mathrm{cfg}}\}. \tag{5}$$

In principle B_{cfg} is an application tolerance: the level past which a configuration's error is unacceptable and its exact size no longer matters. Since the distribution of error scores is skewed and heavy-tailed on several of our datasets, we set B_{cfg} to the 99.9% quantile of s on $\mathcal{D}_{\mathrm{fit}}$ for robustness, rather than to its maximum. At cutoff λ, we send to DFT every test configuration whose predicted error score exceeds λ,

$$S_\lambda = \{R \in \mathcal{D}_{\mathrm{test}} : \hat{s}(R) > \lambda\}, \tag{6}$$

and define the loss of a molecule identity g as the total truncated error of its configurations that are not selected, capped at B,

$$L(\lambda; g) = \min\Big\{ \sum_{R \in \mathcal{R}_g} \ell(R)\, \mathbf{1}\{\hat{s}(R) \le \lambda\},\ B \Big\}. \tag{7}$$

The loss combines the true truncated error $\ell(R)$ with the predicted error score $\hat{s}(R)$, which enters only through the ordering it induces. As a function of λ, $L(\lambda; g)$ is non-decreasing and right-continuous, and it is zero once λ is small enough that every configuration goes to DFT—the conditions required by Angelopoulos et al. (2024) after reversing the sign of λ. CRC additionally requires an upper bound on the loss of a single exchangeable unit. Capping at B inside (7) supplies that bound by construction: $\sup_\lambda L(\lambda; g) \le B$.

The two constants therefore play distinct roles: B_{cfg} bounds the loss of an individual configuration, whereas B bounds the loss of an entire identity. Appendix D details how each is estimated on $\mathcal{D}_{\mathrm{fit}}$. Our goal is to choose $\hat{\lambda}$ on a calibration set $\mathcal{D}_{\mathrm{calib}}$ of n identities so that the expected loss of a test identity g_{n+1} is at most a user-specified level ϵ_0: $\mathbb{E}\big[L(\hat{\lambda}; g_{n+1})\big] \le \epsilon_0$.

Definition 1 (Finite calibrated selection cutoff)**.** Let $\mathcal{D}_{\mathrm{calib}} = \{g_i\}_{i=1}^n$ contain n identities, and define the finite extended set of observed configuration scores

$$\mathcal{H}_n = \{-\infty\} \cup \bigcup_{i=1}^{n} \{\hat{s}(R) : R \in \mathcal{R}_{g_i}\}. \tag{8}$$

Given the budget $\epsilon_0 > 0$, let

$$\mathcal{F}_n = \left\{ \lambda \in \mathcal{H}_n : \frac{\sum_{i=1}^n L(\lambda; g_i) + B}{n+1} \le \epsilon_0 \right\}. \tag{9}$$

The cutoff is

$$\hat{\lambda} = \begin{cases} \max \mathcal{F}_n, & \mathcal{F}_n \neq \varnothing, \\ -\infty, & \mathcal{F}_n = \varnothing. \end{cases} \tag{10}$$

The retention rule is a threshold on $\hat{s}$, so configurations with identical scores are retained or rejected together and only distinct score values are reachable as cutoffs. The term $B/(n+1)$ is the finite-sample correction for the unseen test identity: if $\epsilon_0 < B/(n+1)$ then $\hat{\lambda} = -\infty$, and the whole pool is sent to DFT—the rule is never infeasible, it simply forgoes all savings. The calibrated $\hat{\lambda}$ then controls the risk both for a single test identity and for the test pool as a whole.

Proposition 1 (Risk control per identity)**.** *Let Assumption 1 hold, let $\hat{s}$, B_{cfg} and B be fixed given $\mathcal{D}_{\mathrm{fit}}$. Given the cutoff $\hat{\lambda}$ (10), for a test pool $\mathcal{D}_{\mathrm{test}}$ with $m = |\mathcal{D}_{\mathrm{test}}|$ and a test identity g_{n+1},*

$$\mathbb{E}\big[L(\hat{\lambda}; g_{n+1})\big] \le \epsilon_0, \qquad \mathbb{E}\left[\sum_{g \in \mathcal{D}_{\mathrm{test}}} L(\hat{\lambda}; g)\right] = \sum_{g \in \mathcal{D}_{\mathrm{test}}} \mathbb{E}\big[L(\hat{\lambda}; g)\big] \le m\, \epsilon_0. \tag{11}$$

The proof follows that of Angelopoulos et al. (2024) and we provide the full proof in Appendix C.

We set $\epsilon_0 = \alpha\, \bar{\ell}_{\mathrm{fit}}$, where $\bar{\ell}_{\mathrm{fit}}$ is the mean loss per identity on $\mathcal{D}_{\mathrm{fit}}$. Thus, in expectation, the total error carried by the unselected configurations of a test identity, capped at B, is at most an α fraction of the mean per-identity error of the fitting population. Choosing $\alpha = 0.1$ therefore leaves about a tenth of the error mass in the pool and removes the rest by DFT. Because the bound is anchored to the fitting population rather than the test pool, this does not assert that the realized pool retains at most an α fraction of its own error; Section 3 reports both quantities. The expectation in Proposition 1 is over the exchangeable calibration and test identities, with the fitting stage held fixed. Because $\bar{\ell}_{\mathrm{fit}}$ is fixed by the fitting stage, the bound holds for each α chosen before calibration. It is not a high-probability bound for a single calibration draw, and it is not simultaneous over several values of α (Remark 2).

Table 1: Evaluation datasets. A *configuration* is a geometry and a *molecule identity* is the exchangeable unit; configurations per identity differ by nearly an order of magnitude. MACE-OFF23 contains 228 single-configuration groups without a recorded parent identity.

Dataset	# configs.	# identities	# atoms (mean)	elements
COMP6-GDB07to09-FULL (Smith et al., 2018)	36,000	1,500	7–27 (17.0)	H, C, N, O
MACE-OFF23-FULL (Eastman et al., 2023)	50,195	15,769	3–150 (30.9)	H, C, N, O, F, P, S, Cl, Br, I
QCML-6950-9 (Ganscha et al., 2025)	62,550	6,950	2–23 (11.0)	H, C, N, O, F, P, S, Cl, Br, I

Remark 1 (Role of the predictor)**.** Proposition 1 holds for *any* $\hat{s}$ fixed given $\mathcal{D}_{\text{fit}}$, regardless of the predictor's quality. An uninformative predictor gives $\mathbb{E}[\ell \mathbf{1}\{\hat{s} \le \lambda\}] = \mathbb{E}[\ell]\mathbb{P}(\hat{s} \le \lambda)$, so the budget forces $\mathbb{P}(\hat{s} \le \lambda) \lesssim \epsilon_0/\mathbb{E}[\ell]$ and nearly the entire pool is sent to DFT. Predictor quality therefore governs DFT cost, not validity of risk control; the quantity we are interested in is how many DFT evaluations are saved under fixed budget ϵ_0. The guarantee also holds even if the score is chosen before calibration, using only $\mathcal{D}_{\text{fit}}$ or known facts like whether the user's DFT matches the MLIP's training reference.

3 Experiments

3.1 Setup

We compare our score against four baselines that differ only in the predicted error score $\hat{s}$ used in (7): (i) **Random**, where $\hat{s}$ is uniform and sets the reference cost; (ii) **Frozen raw committee**, where $\hat{s}$ is the rank-combined disagreement among three pretrained MLIP architectures on predicted energy and forces, a standard uncertainty proxy for interatomic potentials (Smith et al., 2018; Schran et al., 2020; Liu et al., 2026); (iii) **Supervised committee**, where $\hat{s}$ is a learned combination of the same three-model disagreement signals, produced by a residual-regression network trained on $\mathcal{D}_{\text{fit}}$; and (iv) **Oracle**, where $\hat{s} = s$ is the true error, a retrospective reference for the cost of perfect ranking rather than a proved lower bound.

All are wrapped in the identical CRC procedure of Section 2.4, so all satisfy the same risk guarantee (further details in Appendix D). Note that each committee requires three pretrained-MLIP evaluations per configuration, whereas ours requires only one MLIP evaluation and the structural descriptor. We report these evaluation counts rather than wall-clock runtimes.

Datasets. We evaluate our method on three molecular datasets that contain DFT labels (Table 1) and three pretrained MLIP models: **MACE-OFF23(M)** (Kovács et al., 2025), **AIMNet2** (Anstine et al., 2025), and **AceFF-2/TensorNet2** (Farr et al., 2026; Simeon & De Fabritiis, 2023). We do not evaluate MACE-OFF23(M) on MACE-OFF23-FULL, whose source it was trained on, leaving 8 dataset–MLIP combinations, or *cells*. Each MLIP is evaluated once per dataset; datasets and MLIPs use different electronic-structure references, so residuals include reference differences as well as model error (Appendix D).

Splits and calibration. We split each dataset by molecule identity into $\mathcal{D}_{\text{fit}}, \mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}}$ at $40/20/40$, so that all configurations of a given identity remain in the same split. The error predictor of Section 2.3 is trained on $\mathcal{D}_{\text{fit}}$ and then frozen, along with energy alignment, size normalization, B_{cfg}, B, and $\bar{\ell}_{\text{fit}}$. For each cell, we draw 20 random $\mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}}$ resamples and sweep $\alpha \in \{0.05, 0.10, \ldots, 0.50\}$, the target fraction of error retained. Proposition 1 converts into the calibration budget $\epsilon_0 = \alpha\, \bar{\ell}_{\text{fit}}$.

Metrics. For each α, we report three quantities: (i) the **DFT calls** $k(\alpha) = |S_{\hat{\lambda}(\alpha)}|$, the number of configurations routed to DFT at that α, reported as a percentage of the test pool; (ii) the **risk ratio** $r(\alpha) = \sum_{g \in \mathcal{D}_{\text{test}}} L(\hat{\lambda}(\alpha); g) \,/\, (|\mathcal{D}_{\text{test}}|\, \epsilon_0)$, which measures how close the realized loss comes to the calibration budget and satisfies $\mathbb{E}[r(\alpha)] \le 1$ by Proposition 1; and (iii) the **realized retained loss** $\rho(\alpha) = \sum_{R \in \mathcal{D}_{\text{test}} \setminus S_{\hat{\lambda}(\alpha)}} \ell(R) \,/\, \sum_{R \in \mathcal{D}_{\text{test}}} \ell(R)$, the error left uncorrected relative to the pool's total error rather than to ϵ_0.

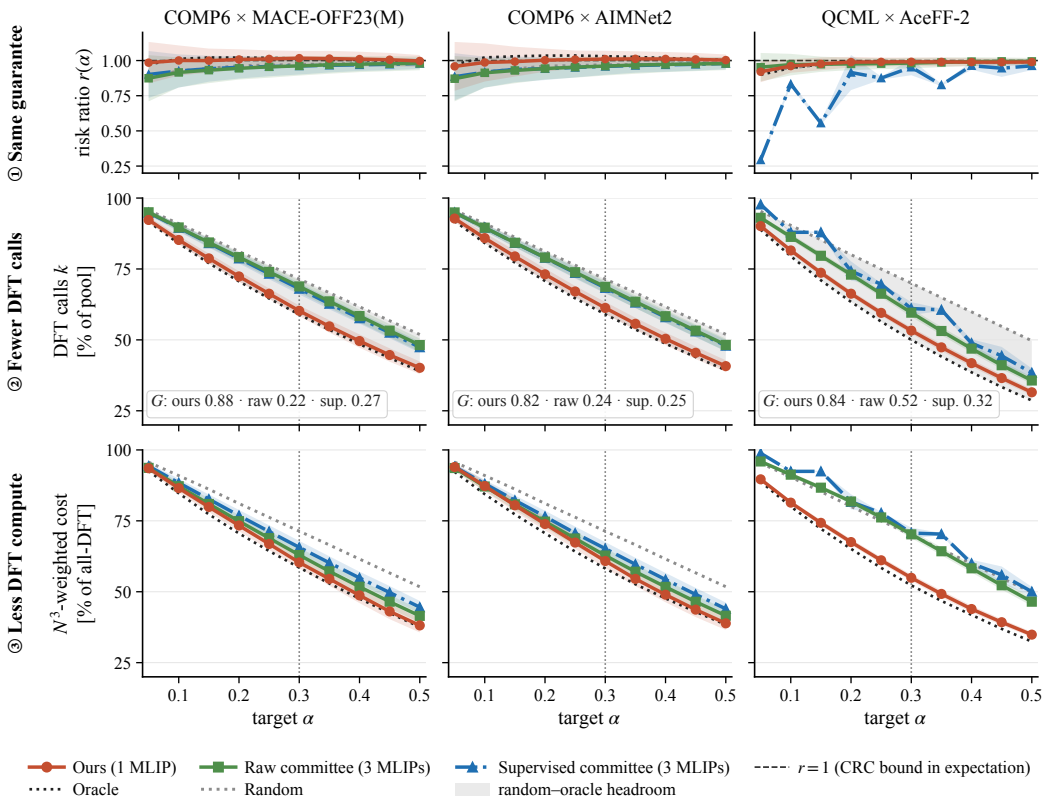

Figure 2: Every score meets the same guarantee; the DFT cost differs. Three representative cells. **Top row**: risk ratio $r(\alpha)$ against the bound $r = 1$. **Middle row**: DFT calls $k(\alpha)$, with the random-to-oracle headroom shaded and each cell's G boxed. **Bottom row**: N^3-weighted DFT cost (% of all-DFT), a size-scaling proxy, not wall time. Lines are means over 20 calibration/test reallocations, and bands are ± 1 SD for ours and the committees. Dotted verticals mark $\alpha = 0.30$.

Since calls weight all configurations equally, we also report the N^3-weighted DFT cost proxy $C_{N^3}(\alpha) = 100 \sum_{R \in S_{\hat{\lambda}(\alpha)}} N(R)^3 \big/ \sum_{R \in \mathcal{D}_{\text{test}}} N(R)^3$ (a size-scaling proxy, not wall time). We also measure the normalized gain,

$$G = \frac{\text{AUC}_{\text{rand}} - \text{AUC}_{\hat{s}}}{\text{AUC}_{\text{rand}} - \text{AUC}_{\text{oracle}}}, \qquad \text{AUC}_{\hat{s}} = \frac{1}{\alpha_{\max} - \alpha_{\min}} \int k(\alpha)\, d\alpha \tag{12}$$

This represents how much of the random-to-oracle headroom is captured by score $\hat{s}$. Note $G = 0$ and $G = 1$ correspond to random and oracle, respectively, and $G < 0$ means it is worse than random selection (anti-selective).

3.2 Results

Same guarantee, but lower DFT cost. Every score passes through the identical CRC rule, so Proposition 1 gives all the same guarantee. With validity fixed by design, models are compared using DFT cost at certified risk. In the top row of Fig. 2, over all cells/scores/targets, the mean risk ratio is at most 1.0359 and the largest single draw is 1.333. Recall that the guarantee is on the expectation, not on each draw. The finite-sample correction $B/(n+1)$ ensures the metric is fair. Without it, the score violates the budget in roughly 52% of calibration draws across $n = 5$–100 identities. But with this correction, violations drop to 0.9% at $n = 10$, trading safety for higher DFT cost when calibration data is scarce (Appendix E). Averaged uniformly over the eight cells (middle row), ours captures $G = 0.81$ of the random-to-oracle headroom, compared to 0.50 and 0.34 for the raw and supervised committees, respectively. At $\alpha = 0.30$, it routes 57.8% of the pool, against 61.9%, 67.5%, 71.6% for raw committee, supervised committee, random, respectively. Our DFT call was comparable with oracle (54.4%). Under N^3 weighting (bottom row) and $\alpha \geq 0.15$, ours is cheaper

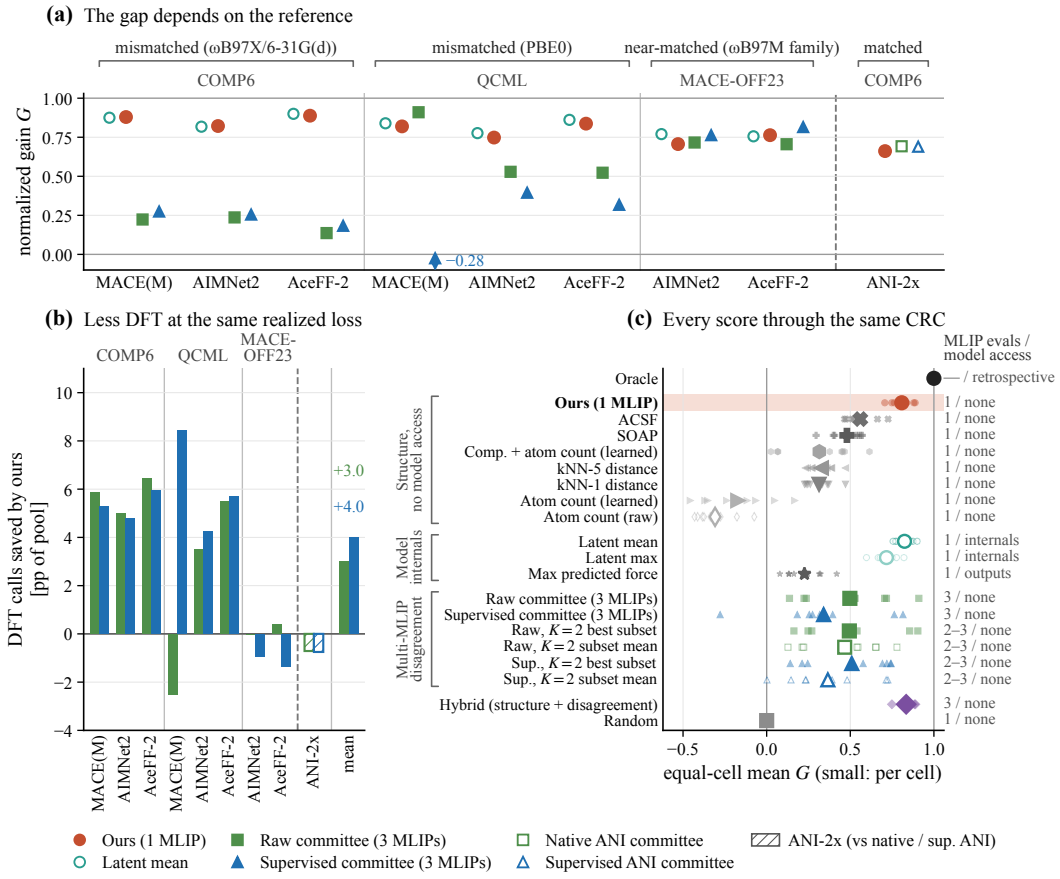

Figure 3: Structure beats disagreement where the MLIPs share error; the gap closes as references converge. (a) Per-cell G. Brackets give the relation between dataset and MLIP training references. Right of the dashed line, ANI-2x × COMP6 is a different MLIP and population, compared with its native and supervised ANI committees. The supervised committee on QCML × MACE-OFF23(M) ($G = -0.28$) is drawn clipped. (b) DFT saved by ours at matched realized loss ρ (committee minus ours, pp of pool), averaged over the common attained ρ range; the mean excludes ANI-2x. (c) Equal-cell mean G (small markers: cells) for every score under the same CRC rule; the right column gives MLIP evaluations per configuration and required model access.

than the raw committee and random in all eight cells and than the supervised committee in seven cells (Appendix F). The realized retained fraction $\rho(\alpha)$ tracks α for every score (Fig. 5, Appendix E). At $\alpha = 0.30$ ours retains $\rho = 0.289$, closest to the target, while the supervised committee (0.251) spends extra DFT on error the user would have tolerated.

Disagreement misses error that the MLIPs share. A residual mixes model error with any mismatch between the MLIP's training reference and the dataset's training reference. In such cases, the three MLIPs share common failure modes. Because they agree on the wrong answer, disagreement-based metrics collectively miss the correct answer. By contrast, our structure-only score captures this shared error from $\mathcal{D}_{\text{fit}}$. Fig. 3(a) shows that on COMP6 (mismatch between DFT–MLIP training references), ours reaches $G = 0.82$–0.89 versus 0.14–0.27 for the committees. On QCML (slightly less of a mismatch), 0.75–0.84 versus 0.32–0.53. One exception is QCML × MACE-OFF23(M), where the raw committee wins (0.91), while the supervised one falls below random (-0.28). We also observed that a few catastrophic, model-specific failures carried nearly all the error (Gini 1.00, against ≤ 0.18 on COMP6). In comparison, on MACE-OFF23 cells (high alignment between DFT–MLIP training references) the supervised committee leads (0.76 and 0.82 versus 0.71 and 0.76). If disagreement misses shared error, committees slightly recover once the training references match. ANI-2x was trained using COMP6's standard reference. There, its own committee reaches $G = 0.69$ and ours reaches 0.66 (whether ANI-2x saw these configurations during training can not be established). The shared error is also not just the standard dispersion

correction. A score built from D3(BJ) alone reaches only $G \leq 0.12$, and adding D3(BJ) to the reference changes our G by at most 0.02 (Appendix G), meaning that our score is not just relearning the dispersion correction. Finally, the committees are not weak by design. Two-member committees average $G \leq 0.51$, and committees built from the target's own model family never beat ours in the five cells where they exist (Appendix H). At matched realized loss ρ, ours still saves 3.0 and 4.0 percentage points of DFT against the raw and supervised committees, positive in 6 of 8 cells each (Fig. 3(b), Appendix F). The reference relation is known before calibration, so the score can be chosen without touching the guarantee (Remark 1). If unsure whether the DFT matches the MLIPs' training reference, the hybrid score (our descriptor plus the three disagreement channels) is always a safe choice. It is at least as good as ours in all eight cells, at the cost of running three MLIPs. This pattern comes from comparing nine settings that are not from a controlled experiment, supporting this explanation. In particular, MACE-OFF23 cells have both a closer reference and larger molecules, so they cannot say which of the two helps the committees.

Structure recovers what model internals know. In Fig. 3(c), ours ($G = 0.81$) is on par with the target MLIP's own representation of 0.82, which needs model access and is fragile to pooling (max pooling: 0.72). It also far exceeds generic descriptors frequently used in chemistry like Atom-Centered Symmetry Functions (ACSF, 0.56) and the Smooth Overlap of Atomic Positions (SOAP, 0.48). Only the latent mean and the hybrid (0.83) are marginally higher, and among scores needing no model internals, ours is strongest. CRC sees a score only through its order, and across all scores G rises with the test Spearman correlation between $\hat{s}$ and s. Three checks show where this ranking comes from (Appendix I). First, no single feature family is essential. Removing any one of them lowers G by at most 0.03. Second, more fitting data helps. G rises from 0.54 with 2.5% of identities to 0.81 with 40%. Third, the pairwise ranking loss is not needed: weights of 0, 0.1, and 1 give nearly the same G (0.807, 0.806, 0.808). As an extension beyond the guarantee, we were curious on what routing does to relative energies (energy differences between geometries of one molecule). At $\alpha = 0.30$, ours cuts their MAE from the all-MLIP level by at least 36.4% in every cell, yet picking each molecule's lowest-energy geometry succeeds less often than with no DFT on the three COMP6 cells. One molecule's MLIP errors tend to share sign and cancel in differences, and replacing some with DFT removes that cancellation (Appendix J). This is a property of partial DFT substitution rather than of any particular score, so a guarantee on energy and force errors does not by itself cover the relative energies. Because Proposition 1 applies to any bounded loss that decreases as the threshold tightens (Remark 3), our framework can certify these objectives directly. The required DFT compute depends on the target: 79–97% for relative-energy errors, but only 8–61% for selection failure (Appendix K).

4 Conclusion

We cast MLIP–DFT routing as risk-controlled selection. A predictor ranks configurations by expected error, and conformal risk control decides where DFT is deployed. Under exchangeability of molecule identities, the expected capped error left to the MLIP stays within a chosen fraction of the fitting-set mean for any score fixed before calibration. Ranking quality therefore changes only DFT cost, and scores can be compared on that cost alone. This comparison matters most when the user's DFT reference differs from the MLIPs' training reference. In this scenario, the MLIPs share much of their error and agree on the wrong answer. Averaged over eight dataset–MLIP cells, a structure-only score fitted on labeled residuals captures 0.81 of the random-to-oracle headroom, against 0.50 and 0.34 for three-MLIP committees, and it matches the target MLIP's own latent features without accessing them. When the references do match, committees recover and demonstrate improved results. The choice of score should therefore follow the reference relation, which is known before calibration. Two results define where the method applies: (i) the fitting and calibration labels are amortized over repeated screening of a population, and (ii) partial DFT replacement can break error cancellation in relative energies, so these objectives should be certified directly, which the framework supports. We provide further limitations and broader impacts in Appendix M and N.

AI USE STATEMENT

In this work, we used generative AI tools to assist in the writing of proofs, provide feedback on research methodology and experiments, implement methods, and interpret results. We have not used generative AI tools to generate synthetic data, develop the theoretical framework, formulate the mathematical claims, propose hypotheses, design the experimental methodology, translate, or perform qualitative data analysis. Additionally, we used generative AI tools to edit software code, draft parts of this paper, produce and refine figures, edit the manuscript for readability, and search for related literature. The research problem, the proposed approach, and all claims are the authors' own: every proof was verified independently of the assistant that proposed it, all AI-assisted code was read and its outputs reproduced, and all references were checked against their sources. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work involves no human subjects, no personally identifiable information, and no new data collection; all datasets and pretrained potentials are publicly released, cited, and used under their licenses (Appendix D). Our method reduces the number of density functional theory (DFT) calculations required at a controlled MLIP annotation risk, and makes downstream molecular screening more accessible to groups without large computing allocations. The main limitations are: Proposition 1 holds in expectation and only under Assumption 1, so a cutoff calibrated on one pool carries no guarantee on configurations from a different distribution, and it controls average risk rather than per-configuration correctness. The method may also be less reliable on datasets with extreme MLIP residuals, since the per-configuration loss is capped at B_{cfg}. We state these limitations in Appendix M. We are aware of no other foreseeable harms, conflicts of interest, or fairness concerns, and we have read and adhere to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

All experiments use public datasets and pretrained potentials, with sources, versions, and preprocessing documented in Appendix D. The structural descriptor is specified in full in Appendix B; the error predictor architecture, training objective, and hyperparameters appear in Section 2.3 and Appendix D; and the calibration procedure appears in Section 2.4, with its proof in Appendix C. The data splits, predictor seeds, and the 20 calibration/test reallocations are described in Appendix D. Code for the selector, the calibration procedure, and all scripts producing the reported tables and figures is included as anonymous supplementary material and will be released publicly upon acceptance.

REFERENCES

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Dylan M. Anstine, Roman Zubatyuk, and Olexandr Isayev. AIMNet2: a neural network potential to meet your neutral, charged, organic, and elemental-organic needs. *Chemical Science*, 16(23): 10228–10244, 2025. doi: 10.1039/d4sc08572h.

Albert P. Bartók, Risi Kondor, and Gábor Csányi. On representing chemical environments. *Physical Review B*, 87(18):184115, 2013. doi: 10.1103/PhysRevB.87.184115.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. In *Advances in Neural Information Processing Systems*, volume 35, pp. 11423–11436, 2022. doi: 10.52202/068431-0830. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/4a36c3c51af11ed9f34615b81edb5bbc-Abstract-Conference.html.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Filippo Bigi, Samuel M. Blau, Vlad Cărare, Michele Ceriotti, Sanggyu Chong, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Edvin Fako, Fabio Falcioni, Andrea C. Ferrari, John L. A. Gardner, Mikołaj J. Gawkowski, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Jonas Grandel, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Cheuk Hin Ho, Stephan Hofmann, Christian Holm, Jad Jaafar, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Panagiotis Kourtis, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Chen Lin, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Louise A. M. Rosset, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Christopher Sutton, Viktor Svahn, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Santiago Vargas, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Thomas Wolf, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry. *The Journal of Chemical Physics*, 163(18):184110, 2025. doi: 10.1063/5.0297006.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535. Article 43.

Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics*, 134(7):074106, 2011. doi: 10.1063/1.3553717.

Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98(14):146401, 2007. doi: 10.1103/PhysRevLett.98.146401.

Filippo Bigi, Sanggyu Chong, Michele Ceriotti, and Federico Grasselli. A prediction rigidity formalism for low-cost uncertainties in trained neural networks. *Machine Learning: Science and Technology*, 5(4):045018, 2024. doi: 10.1088/2632-2153/ad805f.

Jenna A. Bilbrey, Jesun S. Firoz, Mal-Soon Lee, and Sutanay Choudhury. Uncertainty quantification for neural network potential foundation models. *npj Computational Materials*, 11(1):109, 2025. doi: 10.1038/s41524-025-01572-y.

D. R. Bowler and T. Miyazaki. O(N) methods in electronic structure calculations. *Reports on Progress in Physics*, 75(3):036503, 2012. doi: 10.1088/0034-4885/75/3/036503.

Beatriz Cordero, Verónica Gómez, Ana E. Platero-Prats, Marc Revés, Jorge Echeverría, Eduard Cremades, Flavia Barragán, and Santiago Alvarez. Covalent radii revisited. *Dalton Transactions*, (21):2832–2838, 2008. doi: 10.1039/B801115J.

Gábor Csányi, T. Albaret, M. C. Payne, and A. De Vita. "Learn on the fly": A hybrid classical and quantum-mechanical molecular dynamics simulation. *Physical Review Letters*, 93(17):175503, 2004. doi: 10.1103/PhysRevLett.93.175503.

Stefano Curtarolo, Gus L. W. Hart, Marco Buongiorno Nardelli, Natalio Mingo, Stefano Sanvito, and Ohad Levy. The high-throughput highway to computational materials design. *Nature Materials*, 12(3):191–201, 2013. doi: 10.1038/nmat3568.

Bowen Deng, Yunyeong Choi, Peichen Zhong, Janosh Riebesell, Shashwat Anand, Zhuohan Li, KyuJung Jun, Kristin A. Persson, and Gerbrand Ceder. Systematic softening in universal machine

learning interatomic potentials. *npj Computational Materials*, 11(1):9, 2025. doi: 10.1038/s41524-024-01500-6.

Robin Dunn, Larry Wasserman, and Aaditya Ramdas. Distribution-free prediction sets for two-layer hierarchical models. *Journal of the American Statistical Association*, 118(544):2491–2502, 2023. doi: 10.1080/01621459.2022.2060112.

Peter Eastman, Pavan Kumar Behara, David L. Dotson, Raimondas Galvelis, John E. Herr, Josh T. Horton, Yuezhi Mao, John D. Chodera, Benjamin P. Pritchard, Yuanqing Wang, Gianni De Fabritiis, and Thomas E. Markland. SPICE, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10(1):11, 2023. doi: 10.1038/s41597-022-01882-6.

Stephen E. Farr, Stefan Doerr, Antonio Mirarchi, Francesc Sabanes Zariquiey, and Gianni De Fabritiis. Aceff: A state-of-the-art machine learning potential for small molecules, 2026. URL https://arxiv.org/abs/2601.00581.

Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gómez-Bombarelli, and Tommi S. Jaakkola. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=A8pqQipwkt.

Stefan Ganscha, Oliver T. Unke, Daniel Ahlin, Hartmut Maennel, Sergii Kashubin, and Klaus-Robert Müller. The QCML dataset, quantum chemistry reference data from 33.5M DFT and 14.7B semi-empirical calculations. *Scientific Data*, 12:406, 2025. doi: 10.1038/s41597-025-04720-7.

Niklas Geue, Richard E. P. Winpenny, and Perdita E. Barran. Structural characterisation methods for supramolecular chemistry that go beyond crystallography. *Chemical Society Reviews*, 51:8–27, 2022. doi: 10.1039/D0CS01550D.

Yu Gui, Ying Jin, and Zhimei Ren. Conformal alignment: Knowing when to trust foundation models with guarantees. In *Advances in Neural Information Processing Systems*, volume 37, pp. 73884–73919, 2024. doi: 10.52202/079017-2350. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/870ccde24673d3970a680bb48496ed63-Abstract-Conference.html.

Cheuk Hin Ho, Christoph Ortner, and YangShuai Wang. Flexible uncertainty calibration for machine-learned interatomic potentials. *npj Computational Materials*, 12(1):225, 2026. doi: 10.1038/s41524-026-02080-3.

Yuge Hu, Joseph Musielewicz, Zachary W. Ulissi, and Andrew J. Medford. Robust and scalable uncertainty estimation with conformal prediction for machine-learned interatomic potentials. *Machine Learning: Science and Technology*, 3(4):045028, 2022. doi: 10.1088/2632-2153/aca7b1.

Shih-Peng Huang, Nontawat Charoenphakdee, Yuta Tsuboi, Yong-Bin Zhuang, and Wenwen Li. P-DRUM: Post-hoc descriptor-based residual uncertainty modeling for machine learning potentials, 2025. URL https://arxiv.org/abs/2509.02927. NeurIPS 2025 Workshop on Machine Learning and the Physical Sciences (non-archival).

Ying Jin and Emmanuel J. Candès. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24(244):1–41, 2023. URL http://jmlr.org/papers/v24/22-1176.html.

Ryosuke Jinnouchi, Ferenc Karsai, and Georg Kresse. On-the-fly machine learning force field generation: Application to melting points. *Physical Review B*, 100(1):014105, 2019. doi: 10.1103/PhysRevB.100.014105.

Matthias Kellner and Michele Ceriotti. Uncertainty quantification by direct propagation of shallow ensembles. *Machine Learning: Science and Technology*, 5(3):035006, 2024. doi: 10.1088/2632-2153/ad594a.

Walter Kohn and Lu Jeu Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140(4A):A1133–A1138, 1965. doi: 10.1103/PhysRev.140.A1133.

Dávid Péter Kovács, J. Harry Moore, Nicholas J. Browning, Ilyes Batatia, Joshua T. Horton, Yixuan Pu, Venkat Kapil, William C. Witt, Ioan-Bogdan Magdău, Daniel J. Cole, and Gábor Csányi. MACE-OFF: Short-range transferable machine learning force fields for organic molecules. *Journal of the American Chemical Society*, 147(21):17598–17611, 2025. doi: 10.1021/jacs.4c07099.

Salem Lahlou, Moksh Jain, Hadi Nekoei, Victor I. Butoi, Paul Bertin, Jarrid Rector-Brooks, Maksym Korablyov, and Yoshua Bengio. DEUP: Direct epistemic uncertainty prediction. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=eGLdVRvvfQ.

Janice Lan, Aini Palizhati, Muhammed Shuaibi, Brandon M. Wood, Brook Wander, Abhishek Das, Matt Uyttendaele, C. Lawrence Zitnick, and Zachary W. Ulissi. AdsorbML: a leap in efficiency for adsorption energy calculations using generalizable machine learning potentials. *npj Computational Materials*, 9:172, 2023. doi: 10.1038/s41524-023-01121-5.

Zhenwei Li, James R. Kermode, and Alessandro De Vita. Molecular dynamics with on-the-fly machine learning of quantum-mechanical forces. *Physical Review Letters*, 114(9):096405, 2015. doi: 10.1103/PhysRevLett.114.096405.

Kai Liu, Zixiong Wei, Wei Gao, Poulumi Dey, Marcel H. F. Sluiter, and Fei Shuang. Heterogeneous ensemble enables a universal uncertainty metric for atomistic foundation models. *npj Computational Materials*, 12(1):34, 2026. doi: 10.1038/s41524-025-01905-x.

Shams Mehdi, Ilkwon Cho, and Olexandr Isayev. Knowing when to trust machine-learned interatomic potentials, 2026. URL https://arxiv.org/abs/2605.00640.

Felix Musil, Andrea Grisafi, Albert P. Bartók, Christoph Ortner, Gábor Csányi, and Michele Ceriotti. Physics-inspired structural representations for molecules and materials. *Chemical Reviews*, 121 (16):9759–9815, 2021. doi: 10.1021/acs.chemrev.1c00021.

William Overman and Mohsen Bayati. Conformal arbitrage: Risk-controlled balancing of competing objectives in language models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 70305–70337, 2025. doi: 10.52202/085713-2364.

Danny Perez, Aparna P. A. Subramanyam, Ivan Maliyov, and Thomas D. Swinburne. Uncertainty quantification for misspecified machine learned interatomic potentials. *npj Computational Materials*, 11(1):263, 2025. doi: 10.1038/s41524-025-01758-4.

Andrew A. Peterson, Rune Christensen, and Alireza Khorshidi. Addressing uncertainty in atomistic machine learning. *Physical Chemistry Chemical Physics*, 19(18):10978–10985, 2017. doi: 10.1039/C7CP00375G.

Evgeny V. Podryabinkin and Alexander V. Shapeev. Active learning of linearly parametrized interatomic potentials. *Computational Materials Science*, 140:171–180, 2017. doi: 10.1016/j.commatsci.2017.08.031.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Big data meets quantum chemistry approximations: The Δ-machine learning approach. *Journal of Chemical Theory and Computation*, 11(5):2087–2096, 2015. doi: 10.1021/acs.jctc.5b00099.

Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A. Lee, Anubhav Jain, and Kristin A. Persson. A framework to evaluate machine learning crystal stability predictions. *Nature Machine Intelligence*, 7(6):836–847, 2025. doi: 10.1038/s42256-025-01055-1.

Matthias Rupp, Alexandre Tkatchenko, Klaus-Robert Müller, and O. Anatole von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical Review Letters*, 108(5):058301, 2012. doi: 10.1103/PhysRevLett.108.058301.

Christoph Schran, Krystof Brezina, and Ondrej Marsalek. Committee neural network potentials control generalization errors and enable active learning. *The Journal of Chemical Physics*, 153 (10):104105, 2020. doi: 10.1063/5.0016004.

Guillem Simeon and Gianni De Fabritiis. Tensornet: cartesian tensor representations for efficient learning of molecular potentials. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Justin S. Smith, Ben Nebgen, Nicholas Lubbers, Olexandr Isayev, and Adrian E. Roitberg. Less is more: Sampling chemical space with active learning. *The Journal of Chemical Physics*, 148(24): 241733, 2018. doi: 10.1063/1.5023802.

Eli Stavitski and Bert M. Weckhuysen. Infrared and raman imaging of heterogeneous catalysts. *Chemical Society Reviews*, 39:4615–4625, 2010. doi: 10.1039/C0CS00064G.

Oliver T. Unke, Stefan Chmiela, Huziel E. Sauceda, Michael Gastegger, Igor Poltavsky, Kristof T. Schütt, Alexandre Tkatchenko, and Klaus-Robert Müller. Machine learning force fields. *Chemical Reviews*, 121(16):10142–10186, 2021. doi: 10.1021/acs.chemrev.0c01111.

Jonathan Vandermause, Steven B. Torrisi, Simon Batzner, Yu Xie, Lixin Sun, Alexie M. Kolpak, and Boris Kozinsky. On-the-fly active learning of interpretable bayesian force fields for atomistic rare events. *npj Computational Materials*, 6:20, 2020. doi: 10.1038/s41524-020-0283-z.

Joshua A. Vita, Amit Samanta, Fei Zhou, and Vincenzo Lordi. LTAU-FF: Loss trajectory analysis for uncertainty in atomistic force fields. *Machine Learning: Science and Technology*, 6(1):015048, 2025. doi: 10.1088/2632-2153/adb4b9.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, New York, NY, 2005. doi: 10.1007/b106715.

Brandon M. Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John Kitchin, Daniel S. Levine, Kyle Michel, Anuroop Sriram, Taco S. Cohen, Abhishek Das, Sushree Sahoo, Ammar Rizvi, Zachary Ulissi, and Larry Zitnick. UMA: A family of universal models for atoms. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4310. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/bbf23e81b0ad7637fe9a731d0b676ca6-Abstract-Conference.html.

Zixuan Wu, So Won Jeong, Yating Liu, Yeo Jin Jung, and Claire Donnat. Filtering with confidence: When data augmentation meets conformal prediction, 2025. URL https://arxiv.org/abs/2509.21479.

Han Xu, Taoyong Cui, Chenyu Tang, Jinzhe Ma, Dongzhan Zhou, Yuqiang Li, Xiang Gao, Xingao Gong, Wanli Ouyang, Shufei Zhang, and Mao Su. Evidential deep learning for interatomic potentials. *Nature Communications*, 17:937, 2026. doi: 10.1038/s41467-025-67663-y.

Yuzhi Zhang, Haidi Wang, Weijie Chen, Jinzhe Zeng, Linfeng Zhang, Han Wang, and Weinan E. DP-GEN: A concurrent learning platform for the generation of reliable deep learning based potential energy models. *Computer Physics Communications*, 253:107206, 2020. doi: 10.1016/j.cpc.2020.107206.

A Related Works

Configuration-level uncertainty for pretrained MLIPs. Pretrained MLIPs are trained on large and chemically diverse data (Kovács et al., 2025; Batatia et al., 2025; Wood et al., 2025), yet they still fail on individual geometries and retain systematic errors at scale (Fu et al., 2023; Deng et al., 2025). Configuration-level uncertainty is commonly estimated from committee disagreement (Peterson et al., 2017; Schran et al., 2020), Bayesian or evidential models (Vandermause et al., 2020; Xu et al., 2026), or distances in a learned latent space (Hu et al., 2022). For large pretrained models, cheaper estimators avoid training full ensembles (Bilbrey et al., 2025; Bigi et al., 2024; Kellner & Ceriotti, 2024). Liu et al. (2026) combine a heterogeneous ensemble of universal potentials into one uncertainty metric and use an empirically chosen cutoff on it to decide which structures need DFT labels. Such estimates are often overconfident and must be calibrated against labeled errors before use (Perez et al., 2025). Raw and supervised committees represent this lineage as baselines. In our framework, however, the cutoff on any score is calibrated by a CRC framework rather than chosen empirically.

Post-hoc error and reliability prediction for trained MLIPs. There are several methods that share the idea of training a predictor that estimates the error of a pretrained model. DEUP trains a secondary predictor that estimates out-of-sample error (Lahlou et al., 2023), whereas Δ-learning learns a signed correction to a cheap baseline (Ramakrishnan et al., 2015). For MLIPs, recent post-hoc methods draw on the target model's latent representation or training history (Vita et al., 2025; Huang et al., 2025; Mehdi et al., 2026). P-DRUM is closest in target because it regresses energy and force residuals from the descriptors of a trained MACE model (Huang et al., 2025). PROBE instead recasts energy reliability as selective classification on frozen per-atom representations (Mehdi et al., 2026). Our predictor uses a structure-only descriptor whose construction does not depend on the target MLIP (Section 2.2), and a separate predictor is fitted for each dataset–MLIP cell. It needs no model internals but forgoes the model-specific information that a latent representation may carry. Only its ordering enters the routing decision, and the threshold is calibrated separately (Section 2.4).

Uncertainty-guided DFT calls, acquisition, and screening. Learn-on-the-fly molecular dynamics has long coupled quantum-mechanical calculations with a force model that is refitted or extended during the simulation (Csányi et al., 2004; Li et al., 2015). Modern on-the-fly schemes call DFT when a Bayesian error estimate is large and add the result to the training set (Jinnouchi et al., 2019; Vandermause et al., 2020). The DFT result can also replace the unreliable prediction at that step (Jinnouchi et al., 2019). Active-learning pipelines similarly select configurations by committee disagreement or extrapolation grade and retrain on the new labels (Podryabinkin & Shapeev, 2017; Smith et al., 2018; Zhang et al., 2020). Using the DFT value of a flagged configuration is therefore not specific to our setting. What differs is that our target MLIP is never retrained and each deployment pool is routed in a single pass. Our cutoff also carries a finite-sample guarantee on the expected capped retained loss, whereas these cutoffs are set by rule or tuned empirically. Screen-then-verify workflows use machine-learned models to pre-filter candidates for DFT, as in AdsorbML and the Matbench Discovery benchmark (Lan et al., 2023; Riebesell et al., 2025). They select on predicted energy or stability rather than on predicted error. Our selection-success metric is analogous to the AdsorbML success rate (Appendix J).

Conformal calibration and risk control. Conformal prediction calibrates prediction sets or intervals to finite-sample coverage under exchangeability (Vovk et al., 2005; Angelopoulos & Bates, 2023). Risk-controlling prediction sets and Learn-then-Test control more general risks with high probability (Bates et al., 2021; Angelopoulos et al., 2025), while CRC controls the expected value of a bounded monotone loss (Angelopoulos et al., 2024). Related calibration rules select trustworthy foundation-model outputs, filter augmented data, and select candidates with conformal p-values (Gui et al., 2024; Wu et al., 2025; Jin & Candès, 2023). For MLIPs, Hu et al. (2022) calibrate energy intervals with split conformal prediction, and Ho et al. (2026) learn a conformal-inspired scaling of per-atom force uncertainty without a formal coverage guarantee. Both calibrate the uncertainty of individual predictions, whereas our calibration sets a DFT-routing cutoff to control loss of an identity. For our guarantee to hold, we only need exchangeability across molecule identities because each identity contributes to a single bounded loss.

B THE 350-DIMENSIONAL STRUCTURAL DESCRIPTOR

The descriptor is a deterministic function of atomic numbers Z_j and Cartesian coordinates $\mathbf{x}_j$. It uses no reference labels, MLIP predictions, learned embeddings, or fitted parameters. Its nine feature families summarize composition, coordination, distance distortion, close contacts, and local environments (Table 2). These are geometric proxies motivated by chemical structure; they do not compute electronic structure.

B.1 GEOMETRIC QUANTITIES AND SUMMARY CONVENTIONS

The fixed element order is H, C, N, O, F, P, S, Cl, Br, I, with atomic numbers $[1, 6, 7, 8, 9, 15, 16, 17, 35, 53]$. Covalent radii $\rho(Z)$ in Å are $[0.31, 0.76, 0.71, 0.66, 0.57, 1.07, 1.05, 1.02, 1.20, 1.39]$ (Cordero et al., 2008); carbon uses the sp^3 value. Define

$$r_{jk} = \|\mathbf{x}_j - \mathbf{x}_k\|_2, \qquad q_{jk} = \frac{r_{jk}}{\rho(Z_j) + \rho(Z_k)}, \qquad \widetilde{r}_{jk} = \max(r_{jk}, 0.35\,\text{Å}).$$

Table 2: Feature families of the structural vector. The canonical Coordination and PairState blocks contain 60 and 165 coordinates; a fixed compaction removes eight and nine of them, leaving 52 and 156. The descriptor contains the nine families and has 350 dimensions.

Feature family	Dim.	Type	Description
ElementCount	10	Discrete	Global counts of the allowed elements.
Coordination	52	Discrete	Element-resolved proximity-graph coordination states.
PairState	156	Discrete	Element-pair contacts classified as compressed, near nominal separation, or elongated.
ShortRange	21	Continuous	Atomic-number-weighted severity of short interatomic contacts.
Extreme	12	Continuous	Most severe local proximity, strain, and angular-distortion environments.
RBF	18	Continuous	Smooth multiscale pair-distance channels with global and local pooling.
InverseDistance	36	Continuous	Element-independent geometric proximity across several inverse-distance scales.
SoftCrowding	24	Continuous	Smooth filling of local coordination shells over several radii.
ZPotential	21	Continuous	Neighbor-Z-weighted description of the local chemical environment.

The graph G_0 contains all pairs with $q_{jk} \leq 1.20$. A second graph G uses the same candidate edges but restricts H and F to at most one connection, retaining the lowest (q_{jk}, e) where e is the edge identifier used to break distance ties. Coordination and pair-state counts use G_0; local strain and angular summaries use G.

For an atomwise or pairwise vector x, define the ordered summaries

$$\mathcal{S}_7(x) = [\mathrm{sum}, \mathrm{mean}, \mathrm{std}_{\mathrm{pop}}, \max, Q_{0.90}, \mathrm{rms}, \mathrm{count}],$$
$$\mathcal{U}_6(x) = [\mathrm{sum}, \mathrm{mean}, \mathrm{std}_{\mathrm{pop}}, \max, Q_{0.90}, \mathrm{rms}].$$

The operators top1, top2, and top3 sort values in descending order and pad with zeros if fewer than three values exist. On empty edge or pair sets, all summaries and fractions return zero. These conventions fix the dimension for small structures.

B.2 Composition, coordination, and pair states: 218 coordinates

ElementCount records ten element counts. For each element, Coordination counts atoms with G_0 degrees $0, 1, 2, 3, 4, \geq 5$, giving 60 canonical coordinates. For each of the 55 unordered element pairs, PairState counts G_0 edges in three regions: $q < 0.95$, $0.95 \leq q \leq 1.05$, and $1.05 < q \leq 1.20$, giving 165 canonical coordinates. These bins describe distances relative to covalent-radius sums, without assigning bond order or bond energy.

Concatenate the $10 + 60 + 165 = 235$ canonical coordinates and remove the fixed zero-based indices

$$[15, 16, 22, 33, 38, 39, 40, 46, 133, 154, 172, 173, 174, 181, 182, 183, 184].$$

Preserve the order of all remaining coordinates. Eight exclusions belong to Coordination and nine to PairState, leaving $10 + 52 + 156 = 218$ inputs.

B.3 Short-range and extreme local environments: 33 coordinates

For powers $p = 2, 4, 6$, define

$$S_j^{(p)} = \sum_{k \neq j} \frac{Z_j Z_k}{\widetilde{r}_{jk}^{\,p}}.$$

Append $\mathcal{S}_7(S^{(p)})$ for each power, giving 21 ShortRange coordinates; the count entry equals N. Higher powers emphasize closer contacts. Extreme uses three atomwise quantities:

$$R_j = S_j^{(2)}, \qquad W_j = \sum_{\substack{e \in G \\ e \ni j}} \exp\left[-\frac{(q_e - 1)^2}{2(0.20)^2}\right] (q_e - 1)^2,$$

$$B_j = \sum_{\substack{k<\ell \\ k,\ell\in\mathcal{N}_G(j)}} |\sin(\theta_{kj\ell}-\theta_0)|, \qquad \theta_0 = \begin{cases} \pi, & \deg_G(j)=2, \\ 2\pi/3, & \deg_G(j)=3, \\ \arccos(-1/3), & \deg_G(j)\geq 4. \end{cases}$$

Set $B_j = 0$ below degree two. For each of R, W, B, append $[\mathrm{top1}, \mathrm{top2}, \mathrm{top3}, \mathrm{sum}]$, giving 12 coordinates. The angular reference is degree-based, not an orbital hybridization assignment. Its non-monotone sine penalty can vanish for a collapsed degree-two angle; the distance features still record close contacts.

B.4 Radial and inverse-distance summaries: 54 coordinates

For nine ordered centers

$$(\mu_1,\ldots,\mu_9) = (0.568, 0.728, 0.888, 1.048, 1.208, 1.368, 1.528, 1.688, 1.848),$$

define

$$g_\mu(q) = \exp\left[-\tfrac{1}{2}((q-\mu)/0.124)^2\right], \quad P_\mu = \sum_{j<k} g_\mu(q_{jk}), \quad M_\mu = \max_j \sum_{k\neq j} g_\mu(q_{jk}).$$

The RBF block is $[P_{\mu_1}, M_{\mu_1}, \ldots, P_{\mu_9}, M_{\mu_9}]$. For each $p = 1,\ldots,6$, InverseDistance appends $\mathcal{U}_6(\{\widetilde{r}_{jk}^{-p} : j<k\})$, giving 36 coordinates. These unweighted proximity summaries distinguish distance effects from the atomic-number weights in ShortRange.

B.5 Soft crowding and neighboring elements: 45 coordinates

For the ordered scale pairs $(q_0, w) = (0.85, 0.12), (1.00, 0.16), (1.20, 0.22), (1.50, 0.30)$, define

$$C_j^{(q_0,w)} = \sum_{k\neq j} \frac{1}{1+\exp[(q_{jk}-q_0)/w]}.$$

At each scale append $[\mathrm{mean}\, C, \mathrm{std}_{\mathrm{pop}}\, C, \min C, \max C, Q_{0.90}(C)]$. Then append, in order,

$$\mathrm{mean}(C^{0.85,0.12} - C^{1.00,0.16}), \quad \mathrm{mean}(C^{1.00,0.16} - C^{1.20,0.22}),$$
$$\max(C^{0.85,0.12} - C^{1.20,0.22}), \quad \mathrm{std}_{\mathrm{pop}}(C^{0.85,0.12} - C^{1.50,0.30}).$$

These 24 SoftCrowding coordinates use heuristic distance scales, not validated physical coordination-shell boundaries. Finally, for $p = 2, 3, 4$, define $L_j^{(p)} = \sum_{k\neq j} Z_k/\widetilde{r}_{jk}^{\,p}$ and append $\mathcal{S}_7(L^{(p)})$. The resulting 21 ZPotential coordinates retain neighbor element identity without multiplying by the central atomic number. The complete dimension is $218 + 21 + 12 + 18 + 36 + 24 + 21 = 350$.

C Proof of Proposition 1

Proof of Proposition 1. Fix $\mathcal{D}_{\mathrm{fit}}$, $\hat{s}$, target normalization, weights w_t, bounding constants B_{cfg} and B, and ϵ_0. For each λ, the map $g \mapsto L(\lambda; g)$ is a fixed measurable function of a molecule and Assumption 1 makes $\{L(\lambda; g_i)\}_{i\in\mathcal{D}_{\mathrm{calib}}} \cup \{L(\lambda; g_{n+1})\}$ exchangeable. By construction, $L(\lambda; g_i)$ is a non-decreasing function of λ, $0 \leq L(\lambda; g_i) \leq B$, and $L(-\infty; g_i) = 0$.

Let

$$\mathcal{H}_{n+1} = \mathcal{H}_n \cup \{\hat{s}(R) : R \in \mathcal{R}_{g_{n+1}}\}$$

where $\mathcal{H}_n$ follows Eq. 8 and

$$\lambda^\star = \max\Big\{\lambda \in \mathcal{H}_{n+1} : \tfrac{1}{n+1}\textstyle\sum_{i=1}^{n+1} L(\lambda; g_i) \leq \epsilon_0\Big\},$$

which denotes λ in the oracle case when $L(\lambda; g_{n+1})$ is known. The set is non-empty since the condition always holds for $\lambda = -\infty$ so λ^* is well-defined.

For any $\lambda \in \mathcal{F}_n$ we have $\lambda \in \mathcal{H}_{n+1}$ and

$$\frac{1}{n+1}\sum_{i=1}^{n+1} L_i(\lambda) \leq \frac{\sum_{i=1}^n L_i(\lambda) + B}{n+1} \leq \epsilon_0,$$

so $\lambda \le \lambda^\star$; taking the maximum, $\hat{\lambda} \le \lambda^\star$ for $\hat{\lambda}$ defined in Eq. 10 (trivially so if $\mathcal{F}_n = \varnothing$).

Both $\mathcal{H}_{n+1}$ and $\sum_{i=1}^{n+1} L_i$ are invariant under permutations of $g_1, \ldots, g_{n+1}$, so $\lambda^\star$ is a symmetric function of them. Assumption 1 then makes $\mathbb{E}[L(\lambda^\star; g_i)]$ constant in i, whence

$$\mathbb{E}\big[L(\lambda^\star, g_{n+1})\big] = \mathbb{E}\Big[\tfrac{1}{n+1}\textstyle\sum_{i=1}^{n+1} L(\lambda^\star, g_i)\Big] \le \epsilon_0.$$

Since $\hat{\lambda} \le \lambda^\star$ and L_{n+1} is non-decreasing, $L(\hat{\lambda}; g_{n+1}) \le L(\lambda^\star; g_{n+1})$ point-wise, giving $\mathbb{E}[L(\hat{\lambda}; g_{n+1})] \le \epsilon_0$. Each $g \in \mathcal{D}_{\text{test}}$ is exchangeable with $\mathcal{D}_{\text{calib}}$, so this bound applies and linearity of expectation yields $\mathbb{E}[\sum_{g \in \mathcal{D}_{\text{test}}} L(\hat{\lambda}; g)] \le m\epsilon_0$, with no independence required among test molecules. □

Remark 2 (Ties, fallback, and scope)**.** The argument is finite and needs neither continuity of $\hat{s}$ nor absence of ties. When no finite candidate is feasible but $\epsilon_0 \ge B/(n+1)$, $\lambda = -\infty$ still satisfies the criterion; when $\epsilon_0 < B/(n+1)$, the criterion is infeasible, and the rule in Eq. 10 still sends every configuration to DFT. Our method controls *expected* capped identity-level retained loss for an α fixed before calibration and it is worth noting that it does not guarantee $\rho(\alpha) \le \alpha$ for a particular pool and does not bound uncapped physical energy or force error.

Remark 3 (Other bounded losses)**.** The proof only requires that $L(\lambda; g)$ is non-decreasing in λ, bounded by B, zero at $\lambda = -\infty$, and fixed given $\mathcal{D}_{\text{fit}}$. Any per-configuration loss ℓ' with these properties can replace ℓ, and any aggregation over molecules that preserves monotonicity and the bound can replace the capped sum.

D Datasets, experimental protocol, and baselines

We first specify what is fitted, what is frozen, and what calibration is allowed to use. All primary scores are calibrated against the same loss, with the same splits and target grid. Table 1 summarizes the three evaluation datasets.

Splits and resampling. Each dataset is split by its stored molecule identity into $\mathcal{D}_{\text{fit}}, \mathcal{D}_{\text{calib}}, \mathcal{D}_{\text{test}}$ at $40/20/40$. Every fitting-derived quantity and the score predictor are estimated on $\mathcal{D}_{\text{fit}}$ once and frozen. Calibration and test identities are reallocated in 20 random resamples per combination, with $\alpha \in \{0.05, 0.10, \ldots, 0.50\}$. Resample standard deviations measure calibration and pool allocation variability, not training variability. On COMP6 each resample calibrates on 300 identities (7,200 configurations) and evaluates on 600 (14,400); on QCML on 1,390 (12,510) and 2,780 (25,020). On MACE-OFF23, 6,308 fitting groups contain 19,654 configurations with 1–229 configurations per group; calibration and test sets contain 3,154 and 6,307 groups, with test-pool size $20{,}372.8 \pm 333.5$ configurations across resamples. The MACE-OFF23 population has 15,769 groups: 15,541 known parent identities and 228 configurations with unavailable parent identity, each kept as a singleton group. These singletons are not known to be distinct molecules; the exchangeability assumption is stated for the groups as executed. Total atom counts are 612,360 for COMP6, 1,551,806 for MACE-OFF23, and 687,204 for QCML.

Predictor fitting. Each of the three scorers has hidden widths $(256, 256)$, SiLU activations, dropout 0.1, and three Softplus outputs. Training minimizes (4) with AdamW, learning rate 10^{-3}, weight decay 10^{-4}, batch size 256, and at most 300 epochs. Checkpoints use the residual validation L_1 loss with patience 25 and minimum improvement 10^{-5}; the ranking loss is not used for early stopping. A single seed-0 configuration permutation defines the internal 10% validation split of fitting rows for every scorer seed. Feature centering, scaling, and removal of constant coordinates use only the internal training rows, while the residual normalization and loss transforms use the full $\mathcal{D}_{\text{fit}}$. Scorers are trained independently with seeds $0, 1, 2$; each minibatch is paired by one independent uniform permutation (pair-generator seed $100000 + b$ for scorer b), and equal-loss pairs are omitted from the ranking term. The arithmetic mean of the three scores is used directly. The same frozen checkpoints serve all 20 calibration/test allocations, and there is no refit after selection.

Fitting set-derived quantities. The following are estimated on $\mathcal{D}_{\text{fit}}$ and then frozen at calibration: the per-element energy offsets $\hat{a}_z$ that place the MLIP and DFT energies on a common baseline;

the size exponents $\hat{\beta}_t$ and normalizing factors $\hat{c}_t$ for the three channels $t \in \{E, \mathrm{avg}, \max\}$, which together define the error score s; the truncation level B_{cfg}, the 99.9% quantile of s over $\mathcal{D}_{\mathrm{fit}}$, gives the loss $\ell = \min(s, B_{\mathrm{cfg}})$ of Eq. (5). Because molecule identity is the exchangeable unit, the remaining two quantities are calculated over unique identities in $\mathcal{D}_{\mathrm{fit}}$: the bound $B = \max_{g \in \mathcal{I}_{\mathrm{fit}}} \sum_{i \in g} \ell_i$ used in Eq. (9), and the mean per-identity loss $\bar{\ell}_{\mathrm{fit}} = |\mathcal{I}_{\mathrm{fit}}|^{-1} \sum_{g \in \mathcal{I}_{\mathrm{fit}}} \sum_{i \in g} \ell_i$ setting the budget $\epsilon_0 = \alpha\, \bar{\ell}_{\mathrm{fit}}$. Here, $\mathcal{I}_{\mathrm{fit}}$ denotes the set of molecule identities of $\mathcal{D}_{\mathrm{fit}}$. Beyond these, the committees fit inter-model energy offsets to align the scales of different model predictions, and our model trains the score predictor $\hat{s}$, both from $\mathcal{D}_{\mathrm{fit}}$. Random estimates nothing—$\hat{s}$ is a uniform draw—and oracle estimates nothing further, since $\hat{s} = s$. $\mathcal{D}_{\mathrm{calib}}$ is spent entirely on choosing the threshold $\hat{\lambda}$ via Eq. (10).

Per-element energy alignment. A frozen public MLIP checkpoint and a reference DFT dataset do not share an isolated-atom energy convention. Their total energies therefore differ by a fixed amount per atom of each element, and that constant is far larger than the model error we want to measure. The offset of a molecule is the sum of the per-element constants over its atoms, so it is exactly linear in the element counts:

$$e_{\mathrm{true}}(R) - e_{\mathrm{pred}}(R) = \phi(R)^{\top} a + \varepsilon(R), \tag{13}$$

where $\phi(R) \in \mathbb{R}^{K+1}$ holds the counts of the $K = 10$ supported elements (H, C, N, O, F, P, S, Cl, Br, I) together with an intercept, and a collects the per-element offsets a_z and that intercept. Regressing the energy difference on ϕ over $\mathcal{D}_{\mathrm{fit}}$ recovers $\hat{a}$, and $\Delta E(R) = e_{\mathrm{true}}(R) - e_{\mathrm{pred}}(R) - \phi(R)^{\top}\hat{a}$ is the aligned residual that enters the energy channel of s. Because the correction is linear in composition alone, it removes the convention exactly and cannot absorb any geometry-dependent error; it can, however, absorb a composition-linear part of systematic model or reference-level bias, so it does not equate different electronic-structure methods. We fit $\hat{a}$ by iteratively trimmed least squares rather than plain least squares. A handful of MLIP blow-ups, reaching $|\Delta E| \sim 10^8$ eV on QCML, would otherwise dominate the normal equations, and the offset would not be recovered. Writing $\mathrm{TrimLS}_{I_{\mathrm{fit}}}(y)$ for this operator with $y_i = y(R_i)$ and $\phi_i = \phi(R_i)$: initialize $I_0 = I_{\mathrm{fit}}$, the fitting configuration rows, and for $t = 0, \ldots, 5$ compute

$$a^{(t)} \in \arg\min_{a} \sum_{i \in I_t} (y_i - \phi_i^{\top} a)^2, \qquad e_i^{(t)} = y_i - \phi_i^{\top} a^{(t)},$$

$$\tau_t = \max\Big\{ 5\, Q_{0.98}\Big(\{ |e_i^{(t)}| : i \in I_t \} \Big), 10^{-3}\,\mathrm{eV} \Big\}, \qquad J_t = \{ i \in I_{\mathrm{fit}} : |e_i^{(t)}| < \tau_t \}.$$

Return the current coefficients if $|J_t| \le K + 1$ or $t = 5$; otherwise set $I_{t+1} = J_t$ and refit. The threshold quantile uses the current retained rows, but J_t is evaluated over all original fitting rows, so a trimmed row can re-enter. At most six fits are performed.

Target residual normalization. Because the three residual targets in Eq. (2) can scale differently with atom count N, we first remove their systematic size dependence. Let $r_t(R)$ denote the raw residual for target t, with $r_E(R) = |\Delta E(R)|$ the aligned energy residual. For each target separately, the size exponent $\hat{\beta}_t$ is the least-squares slope, with intercept, of $\log r_t(R_i)$ on $\log N(R_i)$ over the fitting rows with $r_t(R_i) > 0$. Writing $v_{it} = r_t(R_i)/N(R_i)^{\hat{\beta}_t}$, the scale is the root mean square over the rows at or below the 0.99 quantile,

$$q_t = Q_{0.99}\big(\{v_{it} : i \in I_{\mathrm{fit}}\}\big), \quad K_t = \{ i \in I_{\mathrm{fit}} : v_{it} \le q_t \}, \quad \hat{c}_t = \left[\frac{1}{|K_t|} \sum_{i \in K_t} v_{it}^2 \right]^{1/2}, \tag{14}$$

where $Q_{0.99}$ is the linearly interpolated empirical quantile with an inclusive boundary. Trimming estimates $\hat{c}_t$ only: it removes no configuration from any set and does not clip u_t. The true and predicted residuals are then normalized using the same frozen quantities,

$$u_t(R) = \frac{r_t(R)}{\hat{c}_t N(R)^{\hat{\beta}_t}}, \qquad \hat{u}_t(R) = \frac{\hat{r}_t(R)}{\hat{c}_t N(R)^{\hat{\beta}_t}}, \tag{15}$$

which removes systematic size dependence and places the three targets on comparable numerical scales. The fitted $\hat{\beta}_t$ and $\hat{c}_t$ are then frozen throughout calibration and inference. The predictor outputs $\hat{u}_t$ directly, so a predicted residual in physical units is recovered by $\hat{r}_t(R) = \hat{c}_t N(R)^{\hat{\beta}_t} \hat{u}_t(R)$.

Committee-based scores. We provide more details on how we calculate $\hat{s}$ for the committee baselines. The committee ranks configurations by disagreement among the three MLIP models used in Section 3. Let $m = 1, \ldots, M$ denote each model ($M = 3$), and let $\hat{e}_m(R)$ and $\hat{\mathbf{f}}_{m,j}(R) \in \mathbb{R}^3$ be model m's predicted total energy and force on configuration R. We use the same frozen public checkpoints for these models as in other experiments in the paper; the committee contains MACE-OFF23(M) as a member on the two MACE-OFF23 cells, where it is excluded as a target.

For each model, raw total energies differ in scale by composition-linear offsets—up to 1.9×10^5 eV per iodine—that would dominate any genuine disagreement. We therefore place every member on a common scale before measuring the spread of predictions. Let $\phi(R) \in \mathbb{R}^{K+1}$ be the element-count vector of R, take member 1 as the reference, and fit

$$\hat{b}_m = \mathrm{TrimLS}_{I_{\mathrm{fit}}}(\hat{e}_1 - \hat{e}_m), \qquad \tilde{e}_m(R) = \hat{e}_m(R) + \phi(R)^\top \hat{b}_m, \tag{16}$$

with $\tilde{e}_1 = \hat{e}_1$, using the same robust trimming as the energy alignment; only model outputs and element counts enter Eq. (16), so the aligned score remains label-free. We then calculate the spread of $\tilde{e}_m(R)$ and $\hat{\mathbf{f}}_{m,j}$. The energy spread is

$$d_E(R) = \Big(\tfrac{1}{M} \sum\nolimits_m \big(\tilde{e}_m(R) - \bar{e}(R)\big)^2 \Big)^{1/2}, \qquad \bar{e} = \tfrac{1}{M} \sum\nolimits_m \tilde{e}_m, \tag{17}$$

and we summarize the atoms two ways to match the two force channels of s:

$$d_{\mathrm{avg}}(R) = \frac{1}{N} \sum_{j=1}^{N} \sigma_j(R), \qquad d_{\max}(R) = \max_{j \in [N]} \sigma_j(R), \tag{18}$$

where

$$\sigma_j(R) = \Big(\tfrac{1}{M} \sum\nolimits_m \big\| \hat{\mathbf{f}}_{m,j}(R) - \bar{\mathbf{f}}_j \big\|_2^2 \Big)^{1/2}, \qquad \bar{\mathbf{f}}_j = \tfrac{1}{M} \sum\nolimits_m \hat{\mathbf{f}}_{m,j}. \tag{19}$$

The three channels carry incomparable units, so the *frozen raw committee* replaces each channel by its rank among the fitting configurations. With n_{fit} fitting configurations,

$$\widehat{F}_{c,\mathrm{fit}}(x) = \frac{1}{n_{\mathrm{fit}}} \sum_{R' \in \mathcal{R}_{\mathrm{fit}}} \mathbf{1}\{d_c(R') \le x\}, \qquad \hat{s}_{\mathrm{comm}}(R) = \sum_c w_c \widehat{F}_{c,\mathrm{fit}}(d_c(R)), \tag{20}$$

with the same weights w that define s in Section 2.3; the inclusive inequality fixes the treatment of ties. The *supervised committee* replaces the fixed rank combination by one residual-regression network with the same hidden widths and output channels as an individual structural scorer, trained with seed 0 on $\mathcal{D}_{\mathrm{fit}}$ from the frozen channels $d_E, d_{\mathrm{avg}}, d_{\max}$ with the L_1 objective. It is a learned-score baseline on the committee's information. Both committees count three pretrained-MLIP evaluations per configuration. *Random* passes a frozen randomized ranking through CRC. *Oracle* passes the true score $\hat{s} = s$ through the same protocol and is retrospective: computing it needs exactly the labels the procedure is meant to avoid, so it is a cost reference, not a deployable selector.

Electronic-structure references and force conventions. The residuals of Eq. (1) compare each MLIP with the dataset's evaluation reference. At least one component of that reference differs from the MLIP's training reference in every combination (Table 3). Residuals therefore include systematic differences between electronic-structure levels as well as MLIP approximation error; the composition-linear alignment removes an offset but does not make functionals, basis sets or dispersion treatments equivalent, and force residuals are not aligned. MACE-OFF23(M) is excluded as a target on the MACE-OFF23 population because of shared training-source provenance (25 exact configurations are recorded as shared with the associated training/validation source); complete training membership is not established for every model, and no blanket unseen or out-of-distribution claim is made. All comparisons use physical forces $\mathbf{F} = -\nabla_R E$ in eV/Å and energies in eV. COMP6's upstream array named `forces` stores the energy gradient, $+\nabla_R E$; ingestion negates that raw vector exactly and converts Hartree/Å to eV/Å with factor 27.211386245988. MACE-OFF23 and QCML references already have physical-force semantics and receive no sign change. Absolute energies are never pooled across different electronic-structure references.

Table 3: Electronic-structure references. Each dataset is paired with the reference used to train each target MLIP; at least one component differs in every combination (Smith et al., 2018; Kovács et al., 2025; Anstine et al., 2025; Farr et al., 2026; Ganscha et al., 2025).

dataset	evaluation reference	target MLIP	MLIP training reference
COMP6	ωB97X/6-31G(d), Gaussian 09	MACE-OFF23(M)	ωB97M-D3(BJ)/def2-TZVPPD
		AIMNet2	ωB97M-D3/def2-TZVPP
		AceFF-2	ωB97M-V/def2-TZVPPD, VV10
QCML	PBE0, FHI-aims	MACE-OFF23(M)	ωB97M-D3(BJ)/def2-TZVPPD
		AIMNet2	ωB97M-D3/def2-TZVPP
		AceFF-2	ωB97M-V/def2-TZVPPD, VV10
MACE-OFF23	ωB97M-D3(BJ)/def2-TZVPPD, PSI4	AIMNet2	ωB97M-D3/def2-TZVPP
		AceFF-2	ωB97M-V/def2-TZVPPD, VV10

Metrics. The deployment DFT percentage and retained fraction ρ are defined in Section 3; their denominator includes every test configuration, and ρ uses the pool's own total truncated loss before routing, whereas the budget uses the frozen fitting-set anchor $\epsilon_0 = \alpha\,\bar{\ell}_{\mathrm{fit}}$. For G in Eq. (12), we first average DFT counts over the 20 resamples at each of the ten α values, then integrate that mean curve by the trapezoidal rule and divide by $\alpha_{\max} - \alpha_{\min} = 0.45$. We calculate one gain per combination and weight the eight combinations equally. The risk ratio

$$r(\alpha) = \frac{\sum_{g\in\mathcal{D}_{\mathrm{test}}} L(\hat{\lambda}(\alpha); g)}{|\mathcal{D}_{\mathrm{test}}|\,\epsilon_0} = \frac{\displaystyle\sum_{g\in\mathcal{D}_{\mathrm{test}}} \min\left\{ \sum_{R\in\mathcal{R}_g} \ell(R)\mathbf{1}\{R \notin S_{\hat{\lambda}(\alpha)}\}, B \right\}}{|\mathcal{D}_{\mathrm{test}}|\,\epsilon_0} \tag{21}$$

applies the cap after summing retained configuration losses within each identity, and its denominator counts every test identity, including those sent entirely to DFT. Proposition 1 gives $\mathbb{E}[r \mid \mathcal{D}_{\mathrm{fit}}] \le 1$ for a target fixed before calibration, not a bound on each resample.

E THE SHARED GUARANTEE

This section extends the top row of Fig. 2 to all eight cells and asks what the finite-sample term of Definition 1 costs. The risk ratio $r(\alpha)$ and the realized retained loss $\rho(\alpha)$ are defined in Section 3.1 and Appendix D.

Across all cells, scores, targets, and reallocations, the largest mean risk ratio over the 20 reallocations is 1.0359 and the largest single draw is 1.333 (Fig. 4); the guarantee bounds an expectation, so individual draws and finite means may exceed 1. At small α the ratio falls well below 1 for every score. The cutoff keeps the calibration loss under $(n{+}1)\epsilon_0 - B$, and $\epsilon_0 = \alpha\,\bar{\ell}_{\mathrm{fit}}$ shrinks with α while the identity bound B does not, so the fixed term $B/(n{+}1)$ consumes most of a small budget. The retained fraction $\rho(\alpha)$ tracks α for every score (Fig. 5); points below the diagonal mean that the rule spent more DFT than the target required.

Fig. 6 holds the ranking score fixed and varies the calibration rule and the number n of calibration identities, at $\alpha = 0.30$. A plug-in cutoff that omits $B/(n{+}1)$ has mean risk/budget 1.01–1.06 for $n = 5$–100 and exceeds the budget in about 52% of calibration draws; a cutoff fixed on $\mathcal{D}_{\mathrm{fit}}$ alone exceeds it in 62%. Finite CRC stays below budget on average (0.17 and 0.9% violations at $n = 10$) but routes the whole pool at $n = 5$ and 74.6% at $n = 100$, against 57.8% at the full calibration size. Violation frequencies describe the observed draws; CRC makes no high-probability claim.

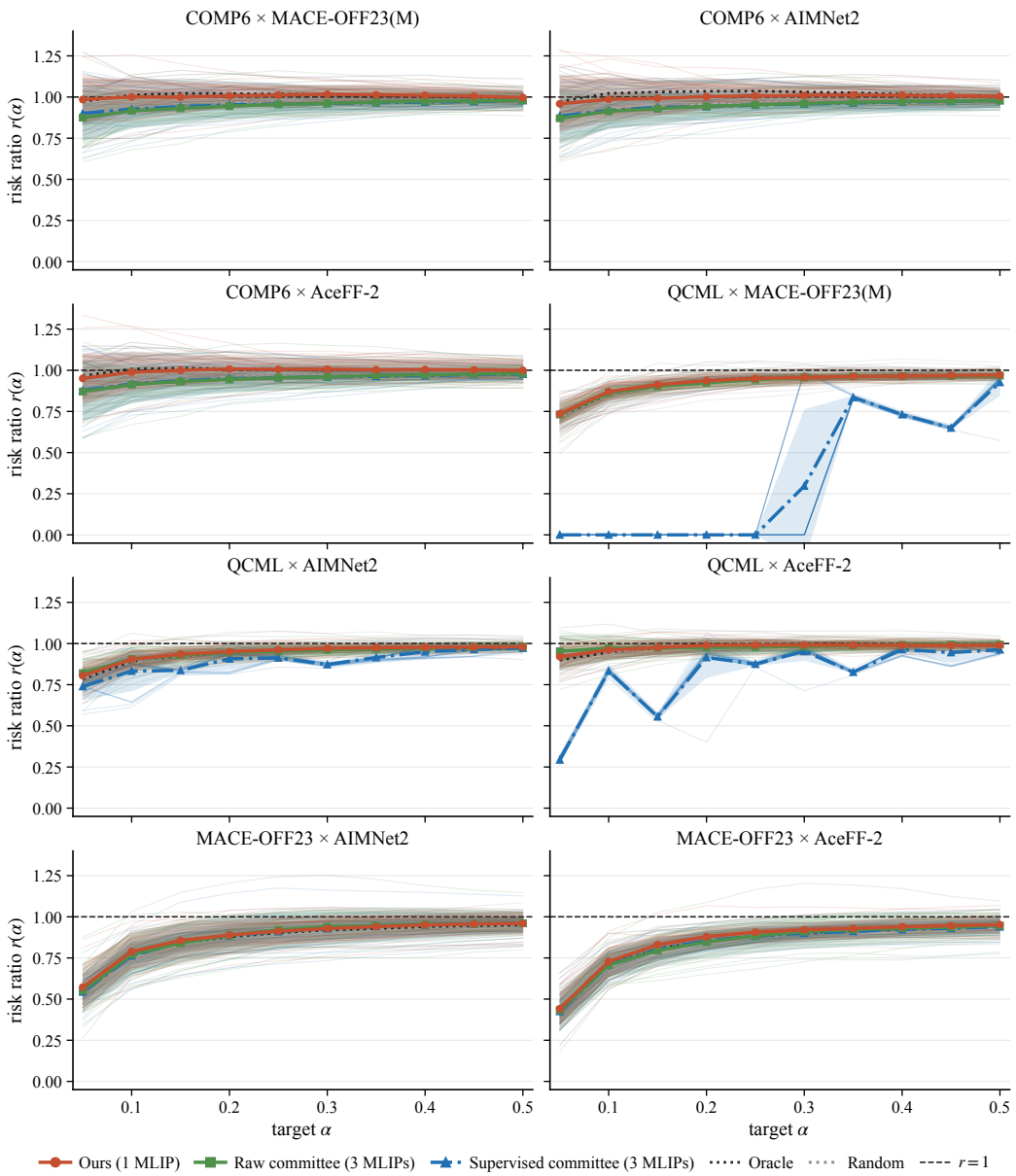

Figure 4: Every score meets the same risk guarantee in every cell. Risk ratio $r(\alpha)$ (realized capped test loss over the budget $\epsilon_0 = \alpha\bar{\ell}_{\text{fit}}$) for all eight cells and five scores. Faint lines are the 20 calibration/test reallocations; bold lines and bands are means ± 1 SD; the dashed line marks $r = 1$. Single draws above 1 are expected because the bound is on the expectation. On QCML × MACE-OFF23(M) the supervised committee routes the whole pool for $\alpha \leq 0.25$, so its r is exactly 0 there.

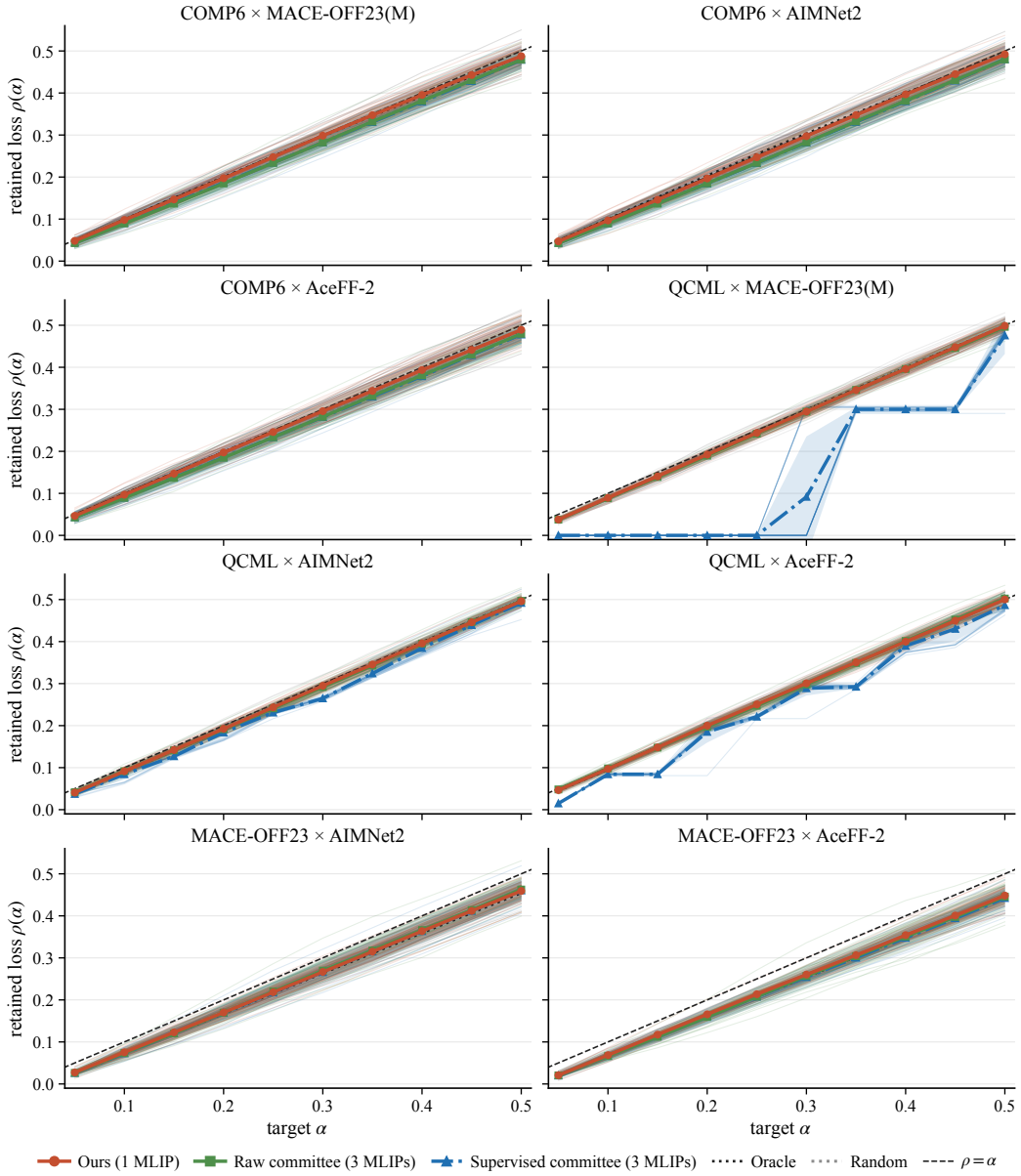

Figure 5: The realized retained loss tracks the target for every score. Retained fraction $\rho(\alpha)$, the capped loss left uncorrected over the pool's own total, for all eight cells; the dashed line is $\rho = \alpha$, a visual reference rather than a guarantee. Faint lines: the 20 reallocations; bold: mean ± 1 SD. Points below the diagonal (MACE-OFF23 cells; the tied supervised committee on QCML) mean that more DFT was spent than the target required.

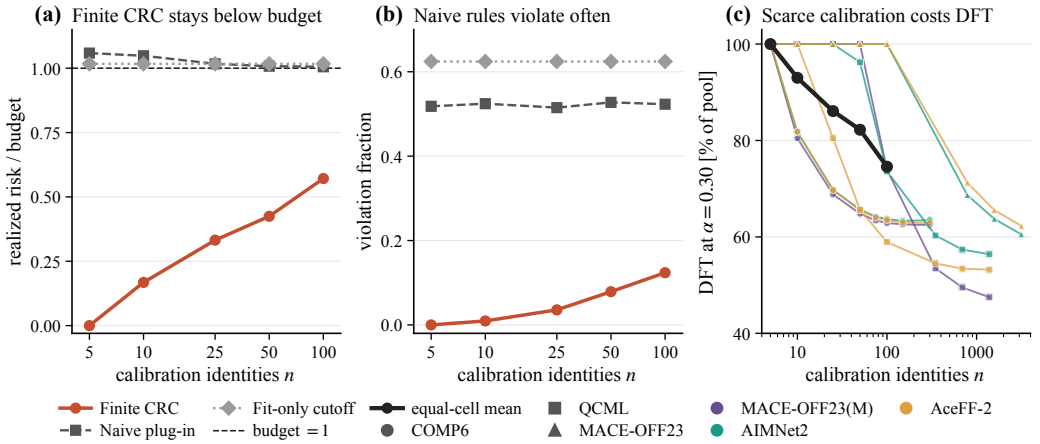

Figure 6: The finite correction keeps the budget; its price is DFT when calibration is scarce. (a) Realized risk/budget and (b) fraction of calibration draws exceeding the budget at $\alpha = 0.30$, for finite CRC, a plug-in cutoff without $B/(n+1)$, and a cutoff fixed on $\mathcal{D}_{\text{fit}}$; equal-cell means over eight cells (200 draws for $n \leq 100$). (c) Deployment DFT at $\alpha = 0.30$: faint lines are cells on their own calibration grids, the bold line the equal-cell mean over the five shared n. The ranking score is fixed throughout; log x axes.

F DFT COST IN EVERY CELL

This section extends the middle and bottom row of Fig. 2 and Fig. 3(b) to every cell. Table 4 gives the equal-cell means, Table 5 the per-cell G, and Table 6 the matched-resource comparisons.

At $\alpha = 0.30$, ours routes less DFT than both committees in 5 of eight cells (Fig. 7); the raw committee is closer to the oracle on QCML × MACE-OFF23(M), and the committees match or beat ours on the two MACE-OFF23 cells. On QCML × MACE-OFF23(M) the supervised committee routes the entire pool for $\alpha \leq 0.25$, and in 14 of 20 reallocations at $\alpha = 0.30$; the rule permits this conservative fallback, and it is why that score has $G < 0$. Weighting each call by N^3 (Fig. 8), ours is cheaper than the raw committee and random in all eight cells for $\alpha \geq 0.15$ and cheaper than the supervised committee in seven, the exception being MACE-OFF23 × AceFF-2. At equal realized loss (Fig. 9), ours saves DFT against each committee in six of eight cells.

Table 4: Main results, equal-weight means over the eight cells. G is the normalized gain Eq. (12); DFT and ρ are the deployment DFT fraction and realized retained loss at $\alpha = 0.30$. The last three columns compare each committee with ours over the common attained support of every cell: DFT saved by ours at matched ρ (Appendix F.1), and the relative-energy MAE and selection-success change of ours at matched deployment DFT (Appendix J).

		at $\alpha = 0.30$			committee vs. ours at matched resources		
Method	MLIP evals	G	DFT [%]	ρ	DFT saved [pp of pool]	rel.-energy MAE [meV]	selection success [pp]
Ours	1	**0.808**	**57.8**	0.289	—	—	—
Raw committee	3	0.497	61.9	0.282	+3.0	−2.0	+0.7
Supervised committee	3	0.341	67.5	0.251	+4.0	−9.7	+1.5
Random	1	0.000	71.6	0.284	—	—	—
Oracle (retrospective)	—	1.000	54.4	0.288	—	—	—

F.1 MATCHED-RESOURCE COMPARISONS

Equal α does not imply equal realized loss: at $\alpha = 0.30$ the supervised committee retains $\rho = 0.251$ against 0.289 for ours (Table 4), so part of its extra DFT buys a lower retained loss. For each cell we therefore form the 20-reallocation mean curves at the ten α values. The common ρ interval I_ρ is the intersection of the ranges attained by ours and the two committees; random, oracle, and the all-MLIP and all-DFT endpoints do not extend it. Within I_ρ, piecewise-linear interpolation defines each curve,

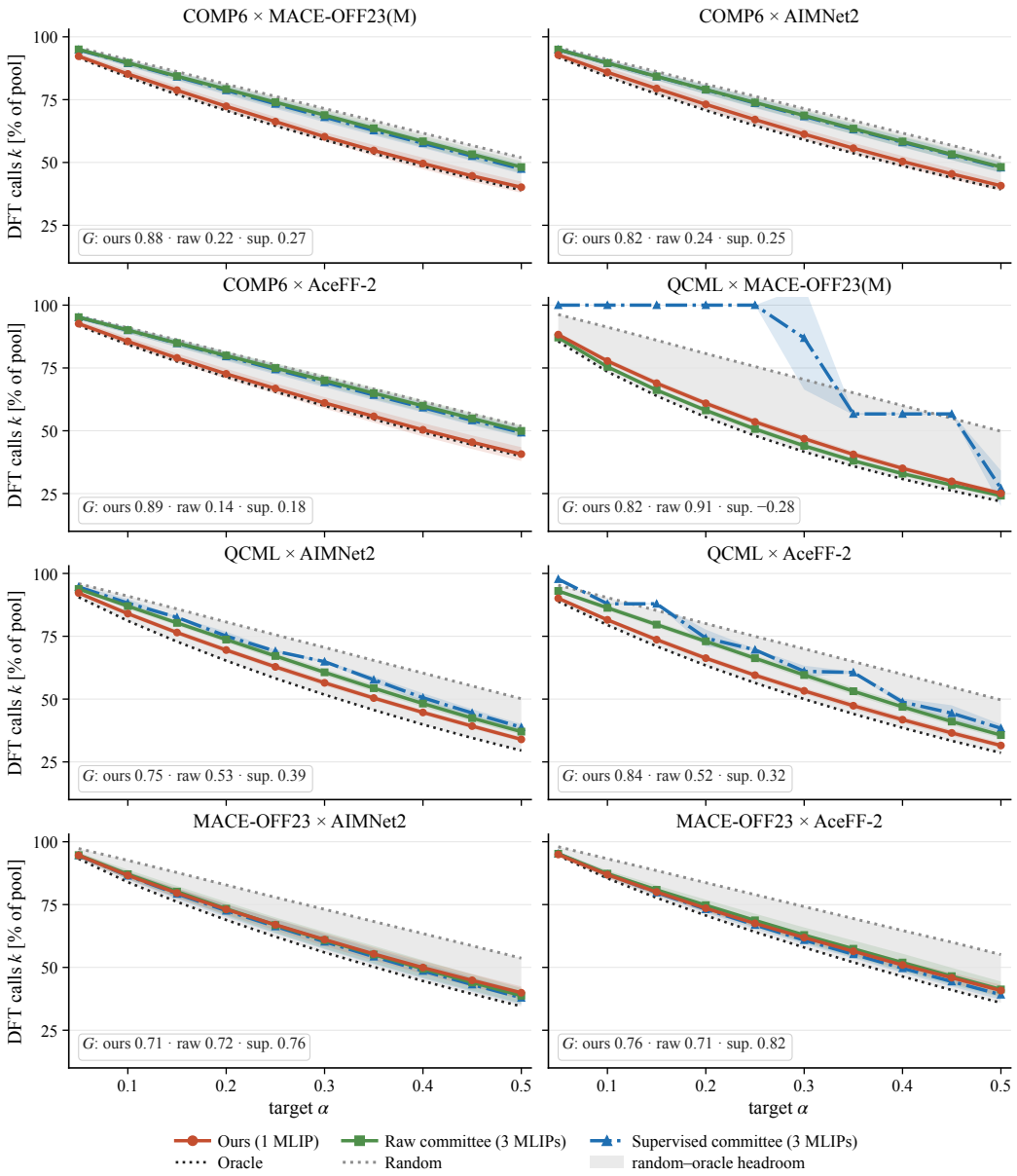

Figure 7: Within the shared guarantee, ours needs the fewest DFT calls of the deployable scores in most cells. DFT calls $k(\alpha)$ for all eight cells; the gray band is the random-to-oracle headroom and each panel prints G for ours and the two committees. Means over 20 reallocations with ± 1 SD bands for ours and the committees. The raw committee wins QCML × MACE-OFF23(M) and the committees match or beat ours on MACE-OFF23. The oracle is retrospective, not deployable.

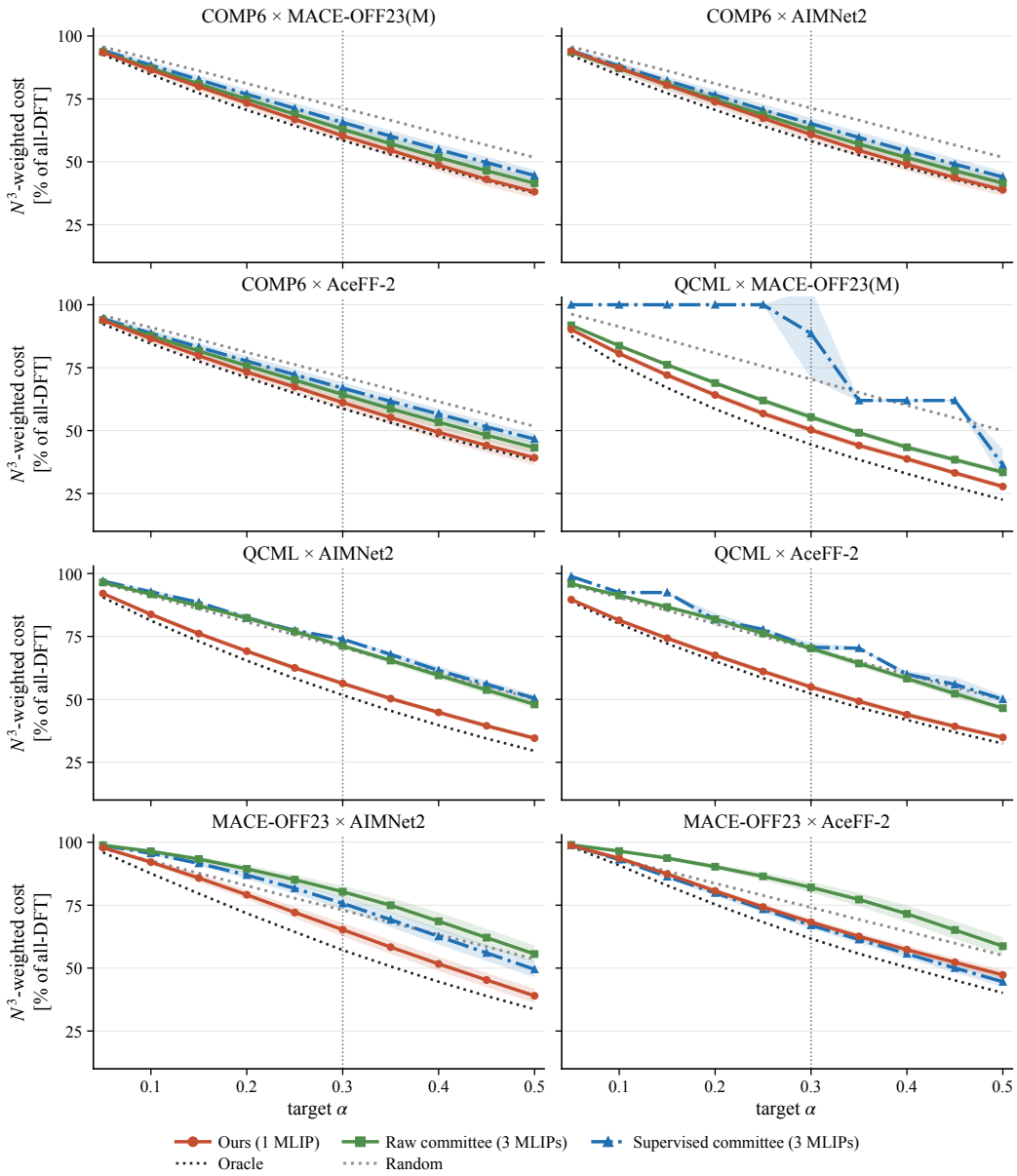

Figure 8: Weighting each DFT call by N^3 preserves, and on QCML widens, the advantage of ours. N^3-weighted DFT cost, the summed cubed atom counts of routed configurations over the same sum for the pool, for all eight cells; the dotted line marks $\alpha = 0.30$. Means over 20 reallocations with ± 1 SD bands for ours and the committees. The proxy captures size scaling, not measured wall time, and the comparison fixes α, not the realized loss.

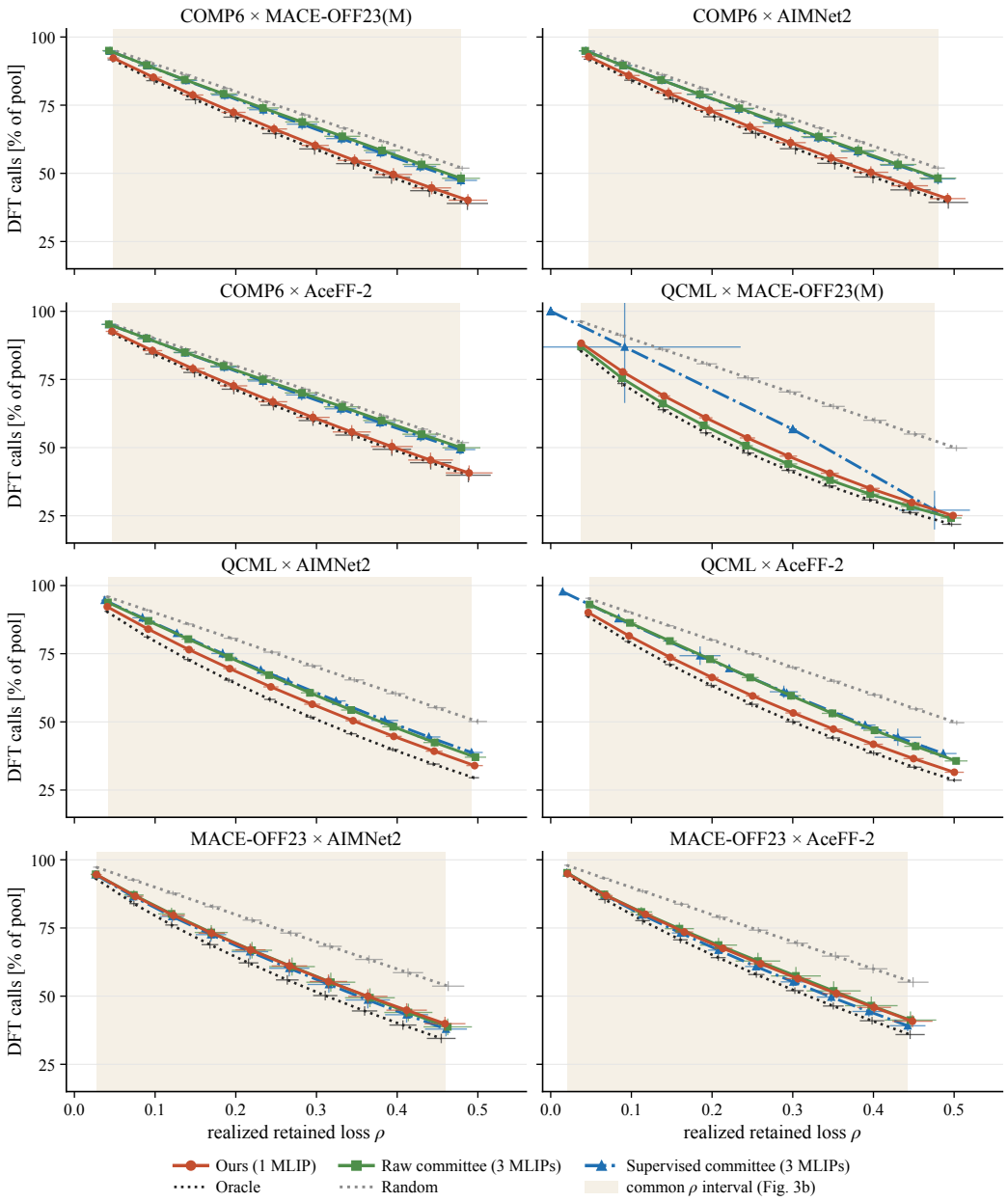

Figure 9: At equal realized loss, ours still calls DFT least in most cells. DFT calls against realized retained loss ρ; each point is one target α at the 20-reallocation mean, with ± 1 SD whiskers in both coordinates. The shaded interval is the common attained ρ range of ours and the two committees, over which Fig. 3(b) averages the DFT difference. On QCML $\times$ MACE-OFF23(M), the supervised committee's first five α values collapse to one all-DFT point at $\rho = 0$.

Table 5: Normalized gain G per cell. Each G uses the 20-reallocation mean DFT-call curve on the α grid. Among ours and the two committees, bold marks the best and underlining the second best; the last three columns are label-free controls (Appendix H), which never exceed ours. The final row is the equal-cell mean.

	primary scores			label-free controls		
Cell	Ours	Raw	Supervised	kNN-1	kNN-5	Max-force
COMP6 × MACE-OFF23(M)	**0.879**	0.223	<u>0.274</u>	0.261	0.267	0.134
COMP6 × AIMNet2	**0.822**	0.237	<u>0.255</u>	0.236	0.244	0.131
COMP6 × AceFF-2	**0.888**	0.136	<u>0.183</u>	0.226	0.232	0.079
QCML × MACE-OFF23(M)	<u>0.820</u>	**0.910**	-0.278	0.362	0.385	0.245
QCML × AIMNet2	**0.748**	<u>0.529</u>	0.395	0.371	0.389	0.322
QCML × AceFF-2	**0.837**	<u>0.523</u>	0.317	0.333	0.361	0.163
MACE-OFF23 × AIMNet2	0.706	<u>0.716</u>	**0.763**	0.471	0.471	0.419
MACE-OFF23 × AceFF-2	<u>0.763</u>	0.705	**0.815**	0.246	0.272	0.315
Mean	**0.808**	<u>0.497</u>	0.341	0.313	0.327	0.226

and for baseline b

$$\overline{\Delta\mathrm{DFT}}_b = \frac{1}{|I_\rho|}\int_{I_\rho} [\mathrm{DFT}_b(\rho) - \mathrm{DFT}_{\mathrm{ours}}(\rho)]\, d\rho, \tag{22}$$

integrated exactly over the union of curve knots, so positive values favor ours. Matched deployment DFT uses the analogous common DFT interval, with differences $\mathrm{MAE}_{\mathrm{ours}} - \mathrm{MAE}_b$ and $\mathrm{Success}_{\mathrm{ours}} - \mathrm{Success}_b$ (Appendix J). Cell values are averaged with equal weight. No extrapolation or endpoint extension is used, and these are descriptive comparisons over common support, not new CRC guarantees.

Table 6: Matched resources over common attained support. (a) Comparisons with the raw and supervised committees; arrows mark the direction favoring ours. DFT saving is baseline minus ours; MAE and success changes are ours minus baseline. The bold mean weights the eight cells equally; pp denotes percentage points. (b) Common intervals used for the comparisons.

(a) Matched-resource comparisons

Dataset	MLIP	DFT saving ↑ [pp]		MAE change ↓ [meV]		Success change ↑ [pp]	
		Raw	Sup.	Raw	Sup.	Raw	Sup.
COMP6	MACE-OFF23(M)	+5.85	+5.28	-16.12	-16.60	+2.19	+2.73
	AIMNet2	+5.01	+4.77	-14.13	-14.40	+2.10	+2.51
	AceFF-2	+6.47	+5.94	-14.63	-15.43	+2.67	+2.91
QCML	MACE-OFF23(M)	-2.51	+8.45	+16.13	-36.87	-1.19	+3.93
	AIMNet2	+3.50	+4.24	+5.56	+3.25	-0.55	-0.32
	AceFF-2	+5.49	+5.69	-1.80	-3.69	+0.34	+0.58
MACE-OFF23	AIMNet2	-0.04	-0.94	+5.40	+4.30	-0.33	-0.21
	AceFF-2	+0.39	-1.34	+3.57	+1.50	+0.15	+0.09
Mean		**+3.02**	**+4.01**	**-2.00**	**-9.74**	**+0.67**	**+1.53**

(b) Common attained support

Dataset	MLIP	Realized loss I_ρ	Deployment DFT I_{DFT} [%]
COMP6	MACE-OFF23(M)	[0.048, 0.479]	[48.17, 92.28]
	AIMNet2	[0.047, 0.480]	[48.22, 92.78]
	AceFF-2	[0.047, 0.478]	[49.95, 92.63]
QCML	MACE-OFF23(M)	[0.038, 0.476]	[27.06, 87.11]
	AIMNet2	[0.042, 0.492]	[38.79, 92.22]
	AceFF-2	[0.048, 0.487]	[38.39, 90.11]
MACE-OFF23	AIMNet2	[0.027, 0.459]	[39.90, 94.57]
	AceFF-2	[0.021, 0.442]	[41.23, 94.94]

G WHY THE GAP OPENS AND CLOSES

This section tests the reading of Fig. 3(a): the structural score wins where the MLIPs share error with one another, and disagreement wins where failures are rare, catastrophic, and model-specific.

Matched reference. ANI-2x is a different MLIP whose eight released members were trained at the nominal COMP6 reference, ωB97X/6-31G(d). We use the mean of members 0–7 as the target, the native committee is their disagreement (population standard deviation), and the supervised ANI committee applies the supervised-committee construction to the same ensemble. If disagreement misses shared error, it should recover here. It does, slightly (Fig. 10): at $\alpha = 0.30$ ours has $G = 0.66$ and routes 57.9%, against 0.69 and 57.2% for both ANI committees and 70.7% for random, and at matched ρ ours needs 0.71 and 0.75 pp more DFT. Training membership of these configurations could not be established, so this is not an unseen-data test.

Isolated dispersion control. D3(BJ) is an empirical correction for long-range dispersion that the evaluation and training references treat differently. A score built from the D3(BJ) term alone reaches $G = 0.05$–0.12 in the six COMP6 and QCML cells, while adding D3(BJ) to the reference and refitting leaves structural G at 0.75–0.90, with each cell changing by at most 0.02 and DFT at $\alpha = 0.30$ by under 1 pp (Fig. 11). This is a diagnostic control, not proof that all electronic-structure mismatch is absent: functional form, basis set, and other effects remain.

Residual concentration. The Gini coefficient of the pool's residual mass is ≤ 0.18 on COMP6, where error is diffuse, and 1.00 on QCML $\times$ MACE-OFF23(M), where a handful of configurations carry essentially all of it (Table 7, Fig. 12). The latter is the one cell where the raw committee beats ours.

Shared and target-specific components. For each configuration, the *consensus* residual is the mean of the three MLIPs' signed aligned energy residuals, or of their per-atom force-residual vectors, and the *target-specific* residual is the target's residual minus that consensus, with force norms taken after vector subtraction. On the resample-0 test pools, the mean Spearman correlation of ours over the eight cells is 0.47/0.54/0.45 with the consensus energy, mean-force, and maximum-force magnitudes, against 0.28/0.45/0.36 with the target-specific ones. The raw committee correlates more with the target-specific force components (0.89/0.86) than with the consensus ones (0.65/0.62). The target-specific part equals the deviation of the target's prediction from the ensemble mean, so the committee's association with it is partly built in; these are descriptive associations, not a causal decomposition.

Model internals and hybrids. The *latent* scores feed the target MLIP's per-atom invariant features (256 dimensions), mean- or max-pooled over atoms, to the same predictor and CRC protocol as ours. The *hybrid* appends the three disagreement channels of the raw committee to our 350-dimensional descriptor. *Native* committees use members of the target's own family: AIMNet2 members 0–3 with member 0 as the frozen target, and MACE-OFF23 S/M/L with M as the target; they are undefined for AceFF-2 targets and for MACE-OFF23(M) on the MACE-OFF23 population. Latent mean pooling (mean $G = 0.82$) exceeds ours in five of eight cells, while max pooling (0.72) drops to 0.60–0.68 on COMP6 (Fig. 13). The hybrid (0.83) is at or above ours in every cell, by $+0.11$ and $+0.07$ on MACE-OFF23 and by at most 0.02 elsewhere; the QCML gains are too small to call improvements. Native committees reach only 0.14–0.24 on COMP6, and the native supervised committee has $G = -1.07$ on QCML $\times$ MACE-OFF23(M). Fig. 14 gives every source in every cell.

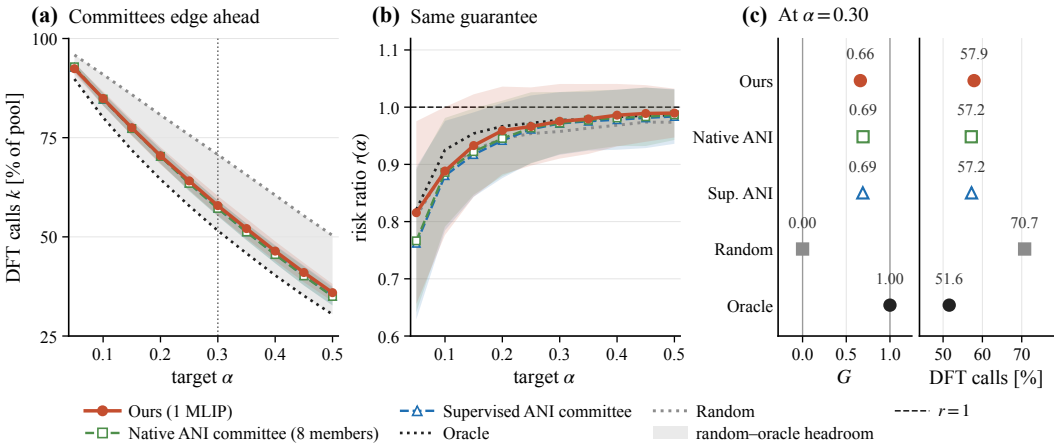

Figure 10: With a matched reference, committees edge ahead, and the structural score stays far above random. ANI-2x × COMP6: (a) DFT calls with the random-to-oracle headroom shaded; (b) risk ratio $r(\alpha)$; (c) G and DFT calls at $\alpha = 0.30$ for ours, the native ANI committee (members 0–7; target is their mean), and the supervised ANI committee. Means over 20 reallocations with ± 1 SD bands for the three selectors. ANI-2x training membership for these configurations is unverified, so this is not an unseen-data test.

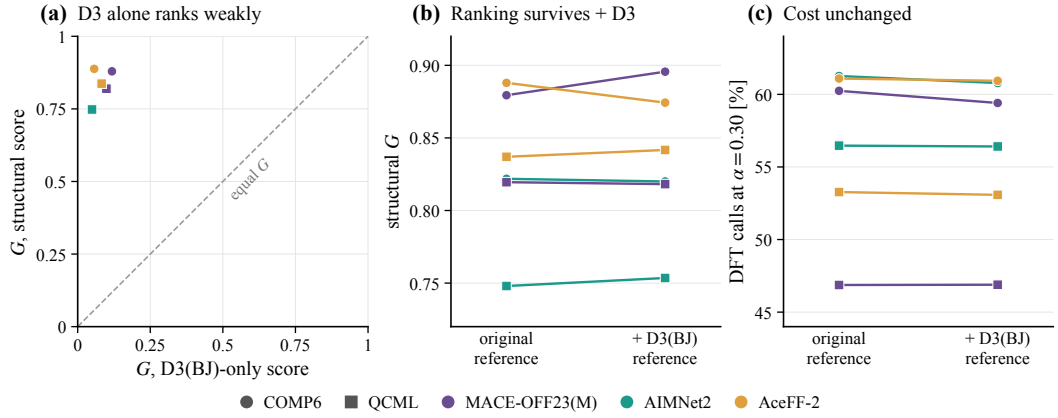

Figure 11: An isolated D3(BJ) term does not explain the structural ranking. (a) Structural G against the G of a D3(BJ)-only score (dashed: equal G). (b) Structural G and (c) DFT calls at $\alpha = 0.30$ with the original reference and after adding D3(BJ) to the reference, joined per cell. Six COMP6 and QCML cells; cell statistics over 20 reallocations. Adjusting the reference changes the loss and refits the predictor, so before and after G use their own random and oracle anchors. A diagnostic, not proof that reference mismatch is absent.

Table 7: Residual concentration and frozen loss bounds. Resample-0 test pool. Gini uses untruncated configuration scores s; "Top 10%" is the untruncated score mass in the highest-mass 10% of identities. $m_{\max}$ is the maximum fitting-identity size, B_{cfg} the configuration truncation level, and B the identity bound of Eq. (7). The last two columns are in meV.

Dataset	MLIP	Gini	Top 10%	$m_{\max}$	B_{cfg}	B	Med. $\lvert\Delta E\rvert$	$p_{99}\lvert\Delta E\rvert$
COMP6	MACE-OFF23(M)	0.18	0.16	24	3.4	45.4	139	516
	AIMNet2	0.16	0.15	24	2.2	46.4	132	474
	AceFF-2	0.16	0.15	24	2.1	47.4	132	443
QCML	MACE-OFF23(M)	1.00	1.00	9	48.2	158.5	485	3,895
	AIMNet2	0.45	0.40	9	34.3	122.6	126	839
	AceFF-2	0.46	0.40	9	11.0	42.4	202	1,151
MACE-OFF23	AIMNet2	0.30	0.42	229	19.3	210.2	41	277
	AceFF-2	0.30	0.45	229	25.0	273.3	82	396

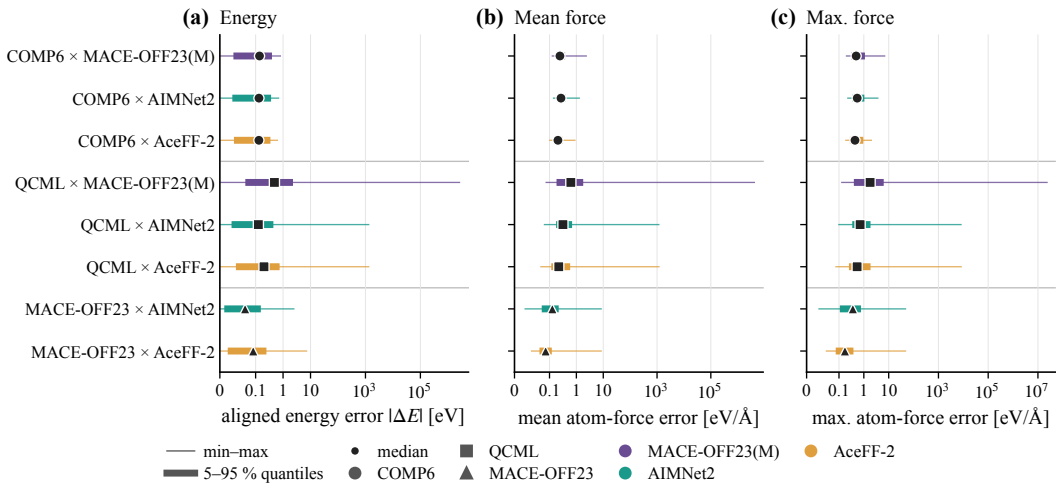

Figure 12: Residual scales differ by orders of magnitude across cells, with catastrophic tails only on QCML. Aligned energy error, mean atom-force error, and maximum atom-force error of the target MLIP on the complete resample-0 test pools. Thin lines span minimum to maximum, thick lines the 5th–95th percentiles, markers the medians. Asinh axes, linear below 0.01, so zero and the largest QCML × MACE-OFF23(M) force error (2.5×10^7 eV/Å) are both visible; nothing is omitted.

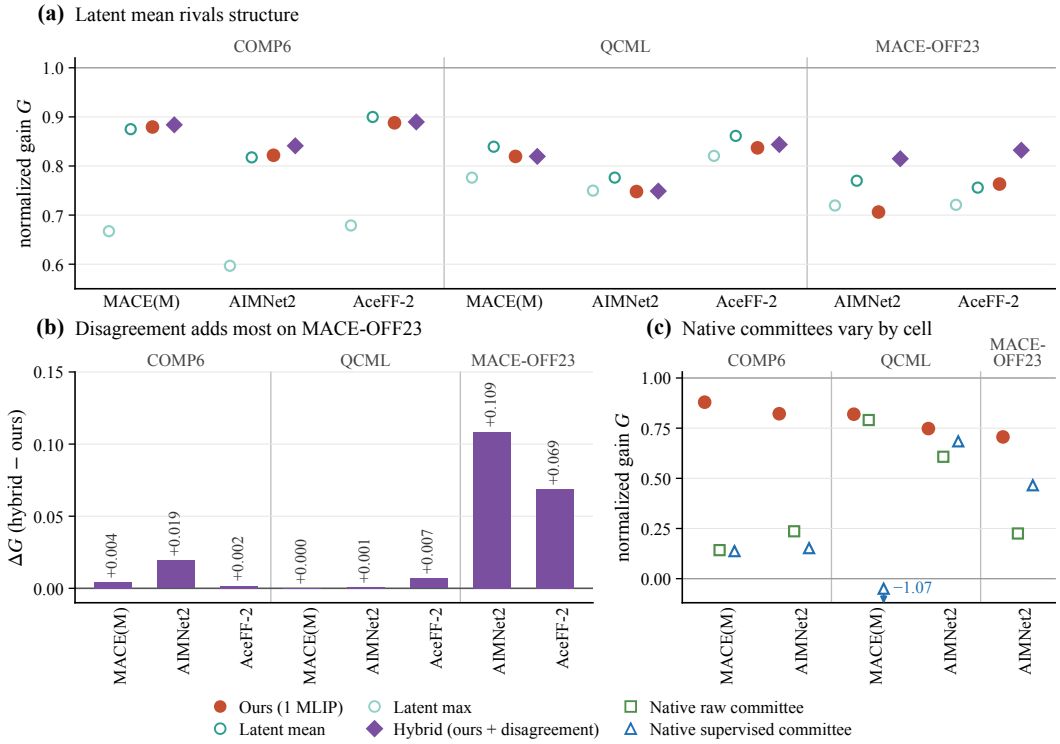

Figure 13: Target-model latents rival but do not dominate the structural score, and disagreement adds most on MACE-OFF23. (a) Per-cell G for ours, latent mean, latent max, and the hybrid (y axis zoomed to 0.55–1). (b) Hybrid minus ours. (c) Native raw and supervised committees against ours in the five cells where they are defined; the native supervised value -1.07 on QCML × MACE-OFF23(M) is drawn clipped. Cell statistics over 20 reallocations. The small QCML hybrid gains are not claimed as improvements.

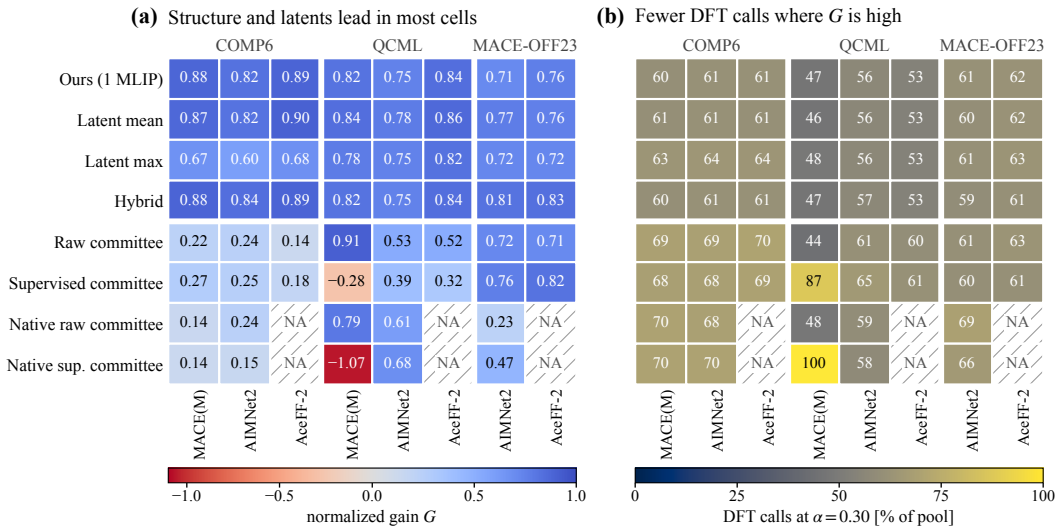

Figure 14: Every score source in every cell. (a) G (diverging map centered at 0) and (b) DFT calls at $\alpha = 0.30$ (fixed 0–100 scale) for ours, latent mean and max, the hybrid, the raw and supervised committees, and the native raw and supervised committees. Cell statistics over 20 reallocations. Hatched NA entries are undefined native committees, not zeros.

H ARE THE COMMITTEES AND CONTROLS STRONG ENOUGH?

This section asks whether the gap in Fig. 3(a, c) reflects weak baselines. Two-member committees keep the target and one other model, or drop the target; the latter still needs three distinct MLIPs, since the target must be evaluated. No two-member subset approaches ours on average (Fig. 15): mean G is 0.445–0.494 for raw and 0.244–0.509 for supervised subsets, against 0.808 for ours. Membership effects are cell-specific; the supervised AIMNet2 + AceFF-2 subset gains $+0.17$ on average, driven by QCML $\times$ MACE-OFF23(M) ($+0.86$), where the full supervised committee is below random. Some subsets beat ours in individual QCML $\times$ MACE-OFF23(M) and MACE-OFF23 cells.

The model-agnostic controls pass through the same loss and CRC rule (Fig. 16): *kNN-1* and *kNN-5*, the descriptor-space distance to the nearest fitting configuration or the mean distance to the nearest five ($G = 0.31$ and 0.33); *max predicted force*, the target MLIP's largest predicted atomic force magnitude (0.23); a *composition + atom count* score and a *learned atom count* score, each trained with our predictor's targets and objective on element counts and atom count, or atom count alone (0.31 and -0.17); and *raw atom count* (-0.31). Ours exceeds every control in every cell. Composition and size carry some signal but do not account for the structural gain.

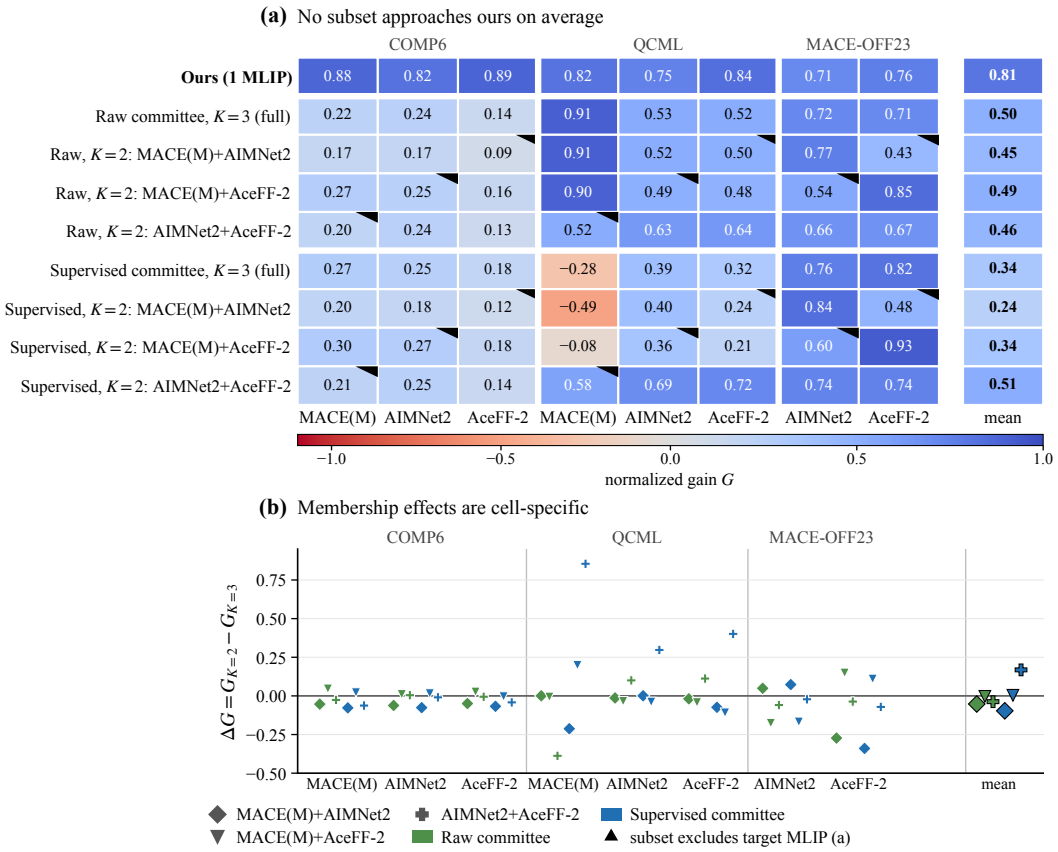

Figure 15: No two-member committee approaches the structural score on average; membership effects are cell-specific. (a) G for ours, the full raw and supervised committees (K=3), and each two-member subset (K=2), per cell and as the equal-cell mean. A black corner triangle marks subsets that exclude the target MLIP; these still need three distinct models, so they are not cheaper than K=3. (b) $\Delta G = G_{K=2} - G_{K=3}$ per cell and subset, raw (green) and supervised (blue), with equal-cell means at right.

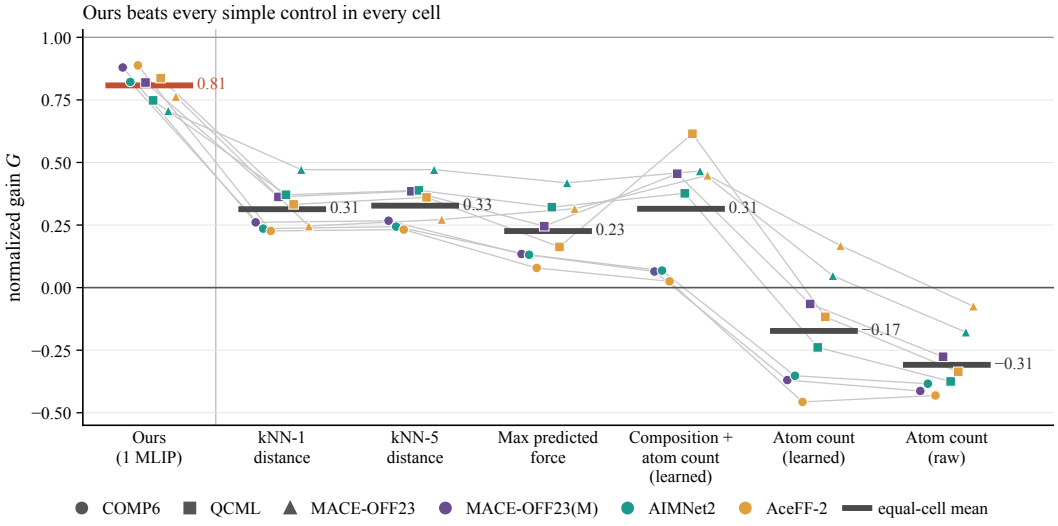

Figure 16: The structural score beats every simple model-agnostic control in every cell. Per-cell G for ours and six controls under the same loss and CRC rule: kNN-1 and kNN-5 descriptor-space distance, the target MLIP's maximum predicted force, learned composition + atom count, learned atom count, and raw atom count. Gray lines join the same cell (marker: dataset; color: MLIP); bars are equal-cell means. Negative G (worse than random) is shown unclipped and does not imply that size is irrelevant.

I WHAT GIVES THE STRUCTURAL SCORE ITS RANKING

This section explains Fig. 3(c). CRC uses a score only through the order it induces, so ranking quality is what becomes DFT savings: across all scores and cells, G rises with the mean test Spearman correlation between $\hat{s}$ and s (Fig. 17). The predictor ranks residuals even where magnitude regression fails, as on QCML × MACE-OFF23(M), where energy R^2 is 0.007 but the Spearman correlation is 0.66 (Fig. 18).

Representation. The full descriptor reaches mean $G = 0.81$ against 0.56 for ACSF and 0.48 for SOAP, two generic atom-centered descriptors trained under the same protocol, and routes 57.8% at $\alpha = 0.30$ against 61.8% and 63.0% (Figs. 19 and 20). This concerns the tested residual-ranking task, not descriptor quality in general. Removing one of the nine feature families changes mean G by at most 0.03 (Coordination, -0.028; PairState, -0.019; RBF, -0.013; the rest within ± 0.006), so the ranking information is spread across families; small positive changes are point estimates, not improvements.

Training choices. More fitting labels improve ranking: mean G rises from 0.54 with 2.5% of population identities to 0.81 with 40% (Fig. 21a). Each fitting size refits the loss quantities as well as the predictor, so each size defines its own proxy loss. The pairwise ranking loss does not materially help at the frozen weight: $\lambda_{\text{rank}} \in \{0, 0.1, 1\}$ give mean $G = 0.807$, 0.806, and 0.808, and $\lambda_{\text{rank}} = 10$ gives 0.70 (Fig. 21b). The frozen $\lambda_{\text{rank}} = 1$ is retained because it was selected before test evaluation.

Fitting-only selection. Architecture, hyperparameters, the three seeds, $\lambda_{\text{rank}} = 1$, and arithmetic averaging were chosen on $\mathcal{D}_{\text{fit}}$ and its fixed internal validation rows by the equal-cell mean Spearman correlation with the capped loss ℓ, then frozen before calibration and testing. The internal split is by configuration, so validation identities also appear in training rows; it supports selection, not an identity-level generalization test. The diagnostics above were computed after freezing and did not enter selection.

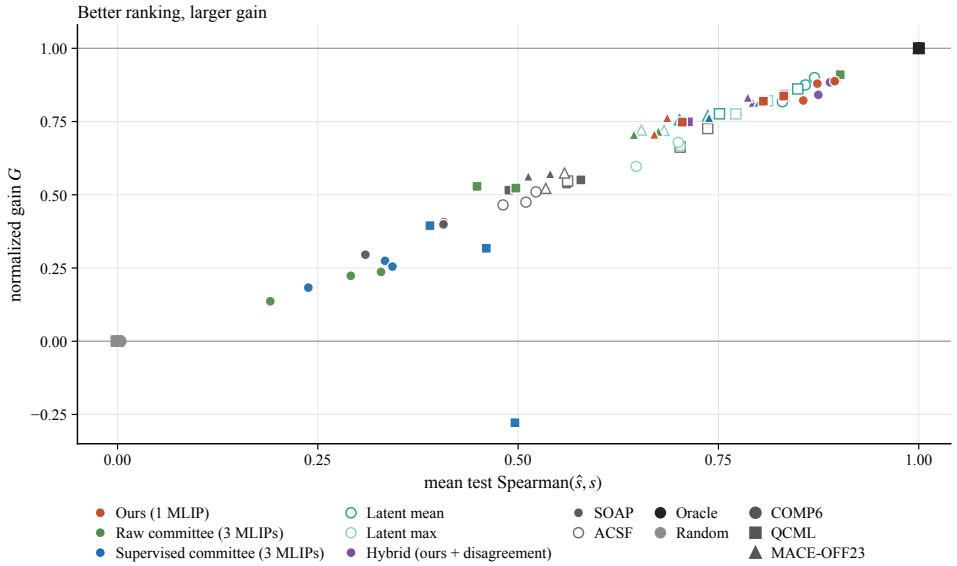

Figure 17: Scores that rank configurations better capture more of the random-to-oracle headroom. G against the mean test Spearman correlation between score and true error, per cell (marker: dataset) and score (color): the five main scores plus latent mean, latent max, hybrid, SOAP, and ACSF (hollow). Spearman values are means over 20 test reallocations; G uses the mean DFT-call curve. Oracle and random sit near $(1, 1)$ and $(0, 0)$.

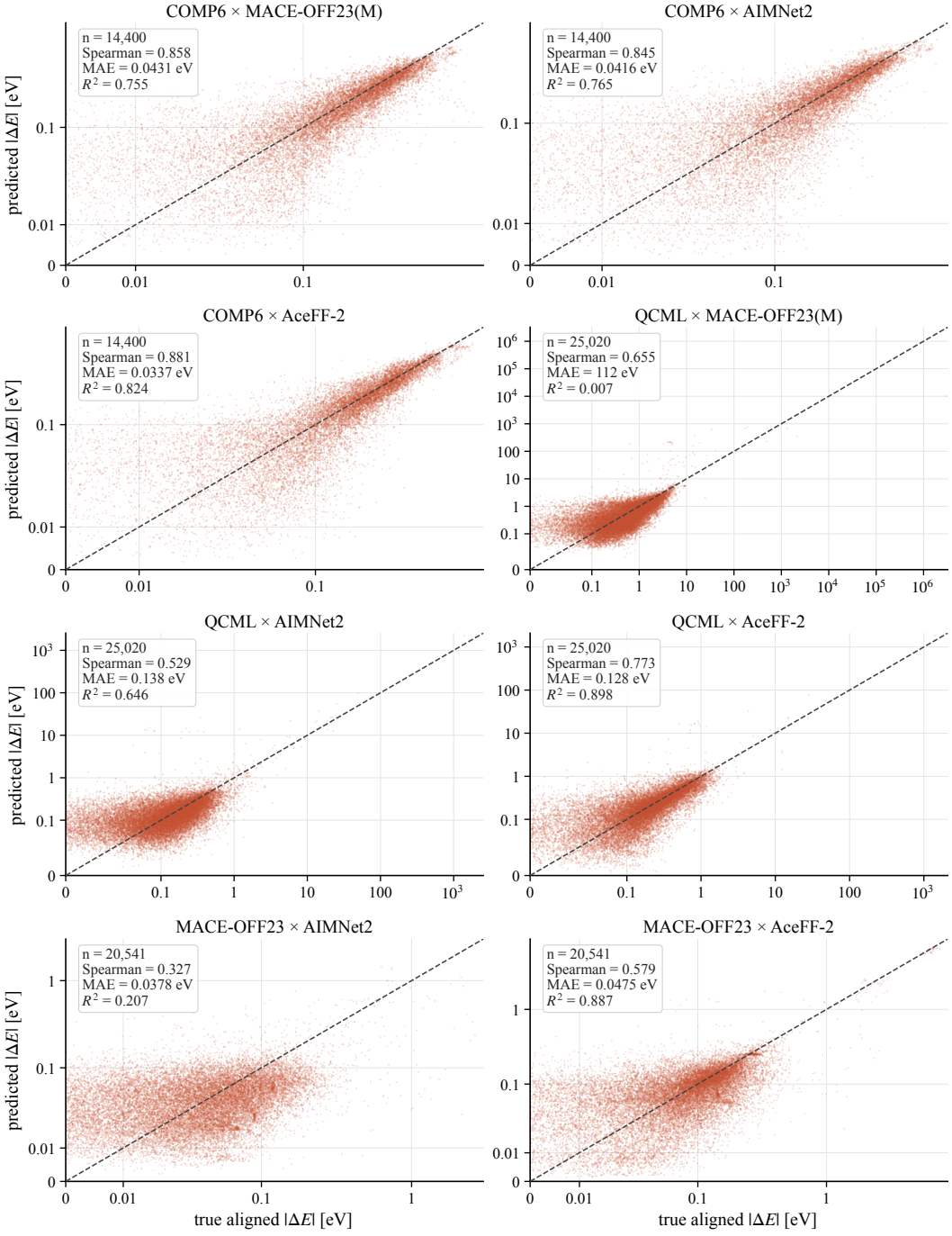

Figure 18: The predictor ranks energy residuals even where it misestimates their size. True against predicted aligned $|\Delta E|$ (energy channel of the frozen predictor) for every resample-0 test configuration; asinh axes, linear below 0.01 eV; dashed line $y = x$. Boxes give n, Spearman correlation, MAE, and R^2 on raw residuals with outliers kept. The QCML × MACE-OFF23(M) axis reaches about 10^6 eV because of a few catastrophic residuals, which is also why its R^2 is near 0.

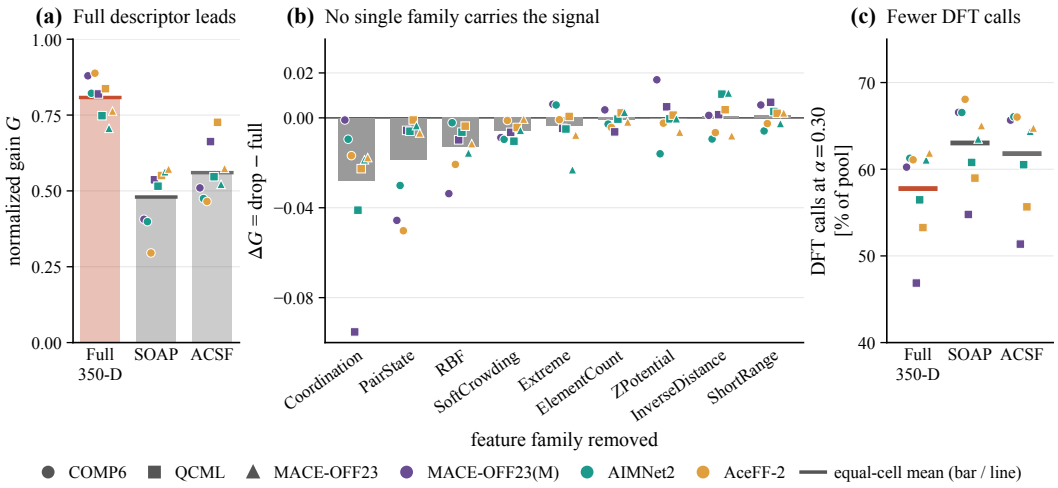

Figure 19: The full descriptor outperforms SOAP and ACSF, and no single feature family carries the signal. (a) Equal-cell mean G (bars) and per-cell values (dots) for the full 350-dimensional descriptor, SOAP, and ACSF. (b) ΔG (drop minus full) for each single-family drop, ordered by mean effect. (c) DFT calls at $\alpha = 0.30$ (y axis zoomed to 40–75%). All use the frozen fitting protocol and 20 reallocations; small positive ΔG values are not significant improvements.

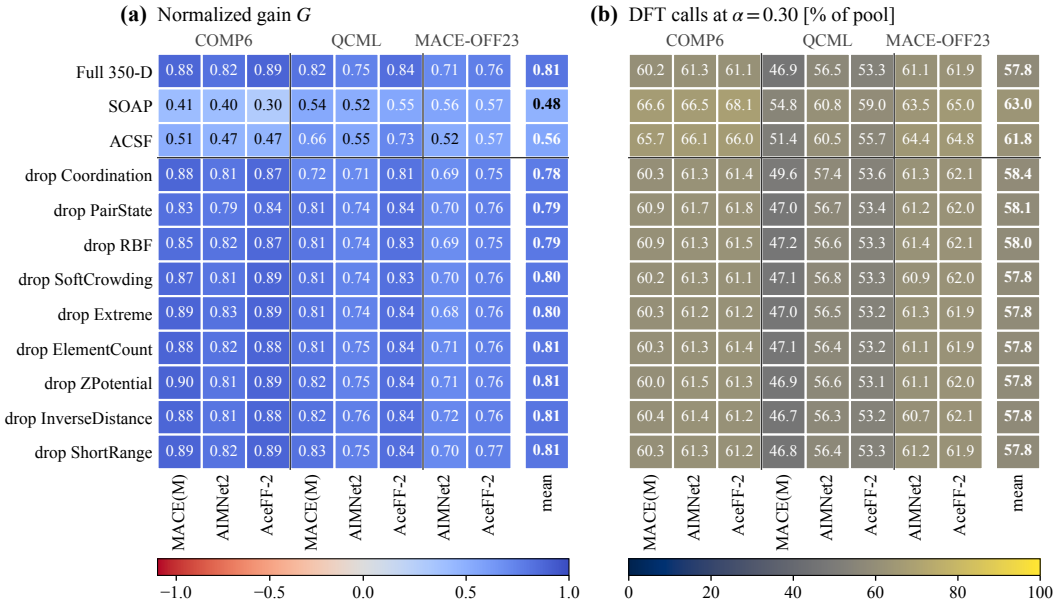

Figure 20: Descriptor effects are consistent across cells. (a) G (diverging scale centered at 0) and (b) DFT calls at $\alpha = 0.30$ (fixed 0–100 scale) for every representation and single-family drop in every cell; the last column is the equal-cell mean. Rows below the rule are drops from the full descriptor, ordered by mean ΔG.

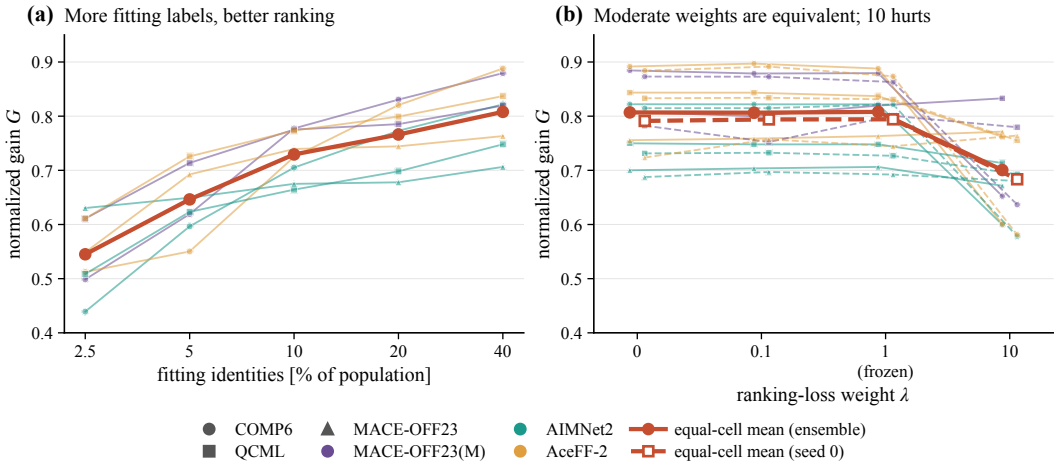

Figure 21: More fitting labels improve ranking; the ranking-loss weight matters only when large. (a) G against fitting-set size (2.5–40% of population identities, nested subsets of the original fitting pool), one thin line per cell and the bold equal-cell mean. (b) G against the ranking-loss weight $\lambda \in \{0, 0.1, 1 \text{ (frozen)}, 10\}$ for the three-seed ensemble (solid) and seed 0 (dashed). Lines join cell values and are not uncertainty intervals.

J What routing does to chemical observables

This section extends the relative-energy results of Section 3.2. These quantities are outside Proposition 1.

Definitions. *Deployment DFT* is the percentage of the test pool routed to DFT. *Hybrid energies* take the reference energy for routed configurations and the MLIP energy, with the frozen fitting-only alignment, otherwise; the all-MLIP endpoint uses raw predictions, since a composition-linear offset cancels within a molecule. For each test identity with at least two configurations, *relative-energy MAE* averages the absolute errors of all unordered energy differences, then averages identities equally. *Selection success* is the fraction of identities whose lowest hybrid-energy configuration lies within 1 kcal/mol (43.3641 meV) of the identity's reference minimum. All 600 COMP6 and 2,780 QCML test identities qualify; MACE-OFF23 eligibility varies across reallocations and excludes singleton groups.

Findings. At matched deployment DFT (Fig. 22; procedure in Appendix F.1), ours lowers relative-energy MAE by 2.0 and 9.7 meV against the raw and supervised committees, better in 4 and 5 of eight cells; the supervised mean is dominated by QCML × MACE-OFF23(M) (−36.9 meV). Selection success rises by 0.7 and 1.5 pp, better in 5 and 6 of eight. The COMP6 curves separate clearly, while QCML and MACE-OFF23 largely overlap (Fig. 23). At $\alpha = 0.30$, ours reduces relative-energy MAE from the all-MLIP endpoint by at least 36.4% in every cell, yet selection success falls below all-MLIP on the three COMP6 cells. Routing removes 73–99.7% of the full-pool mean energy error and 65–99.8% of the mean force error, but rare large force errors remain among retained configurations, up to 50.300 eV/Å on MACE-OFF23 × AceFF-2 (Fig. 24). Truncation at B_{cfg} affects at most 0.159% of configurations but removes 97.116% of the raw score mass on QCML × MACE-OFF23(M); CRC controls only the capped loss.

J.1 Signed-residual cancellation

The MLIP residuals of one molecule's configurations often share sign, and same-sign residuals cancel in an energy difference. Routing one configuration to DFT sets its residual to zero and removes that cancellation. Before routing, the identity-balanced correlation of paired residuals is 0.92–0.96 on COMP6 and MACE-OFF23 × AceFF-2, and routing with ours at $\alpha = 0.30$ lowers it to 0.55–0.60 (Fig. 25). On QCML the all-MLIP correlation is near zero and rises after routing because the catastrophic residuals are removed. As a result, a molecule's relative-energy loss need not fall monotonically as DFT is added: 45–95% of eligible identities have a non-monotone routing path

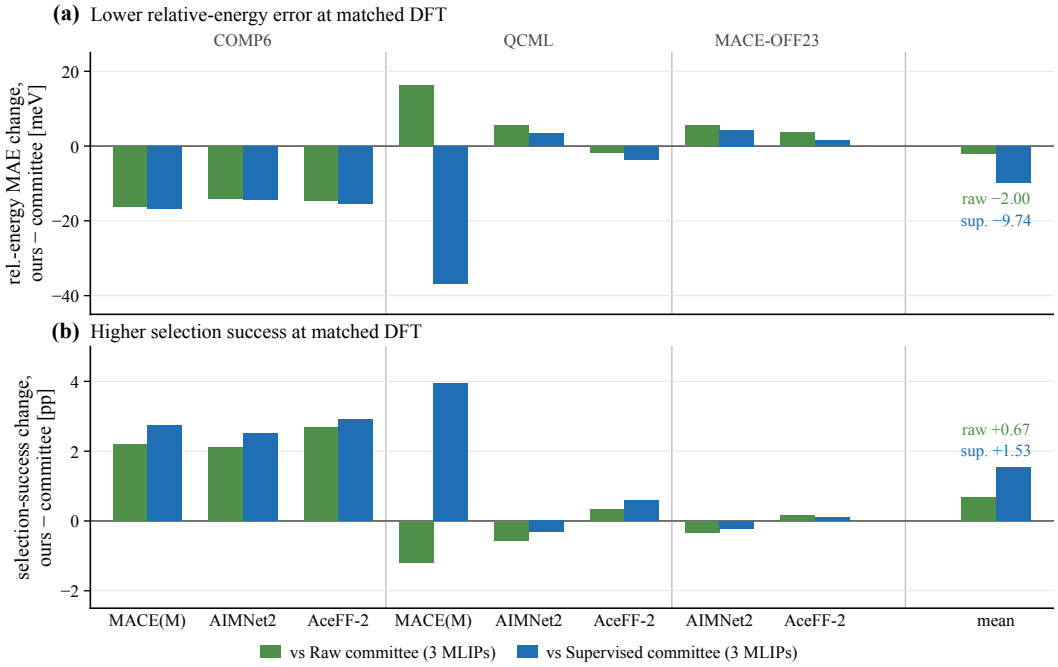

Figure 22: At matched deployment DFT, ours lowers relative-energy error and raises selection success on average, but not in every cell. Per-cell difference, ours minus committee, against the raw (green) and supervised (blue) committees: (a) relative-energy MAE (lower is better); (b) selection success (higher is better). The right group is the equal-cell mean. Each value averages the difference of the 20-reallocation mean curves over the cell's common attained deployment-DFT interval. Descriptive, not a new guarantee.

for relative-energy MAE and 1–46% for selection failure (Fig. 26). These aggregate diagnostics are consistent with lost cancellation but do not identify the cause of each selection failure.

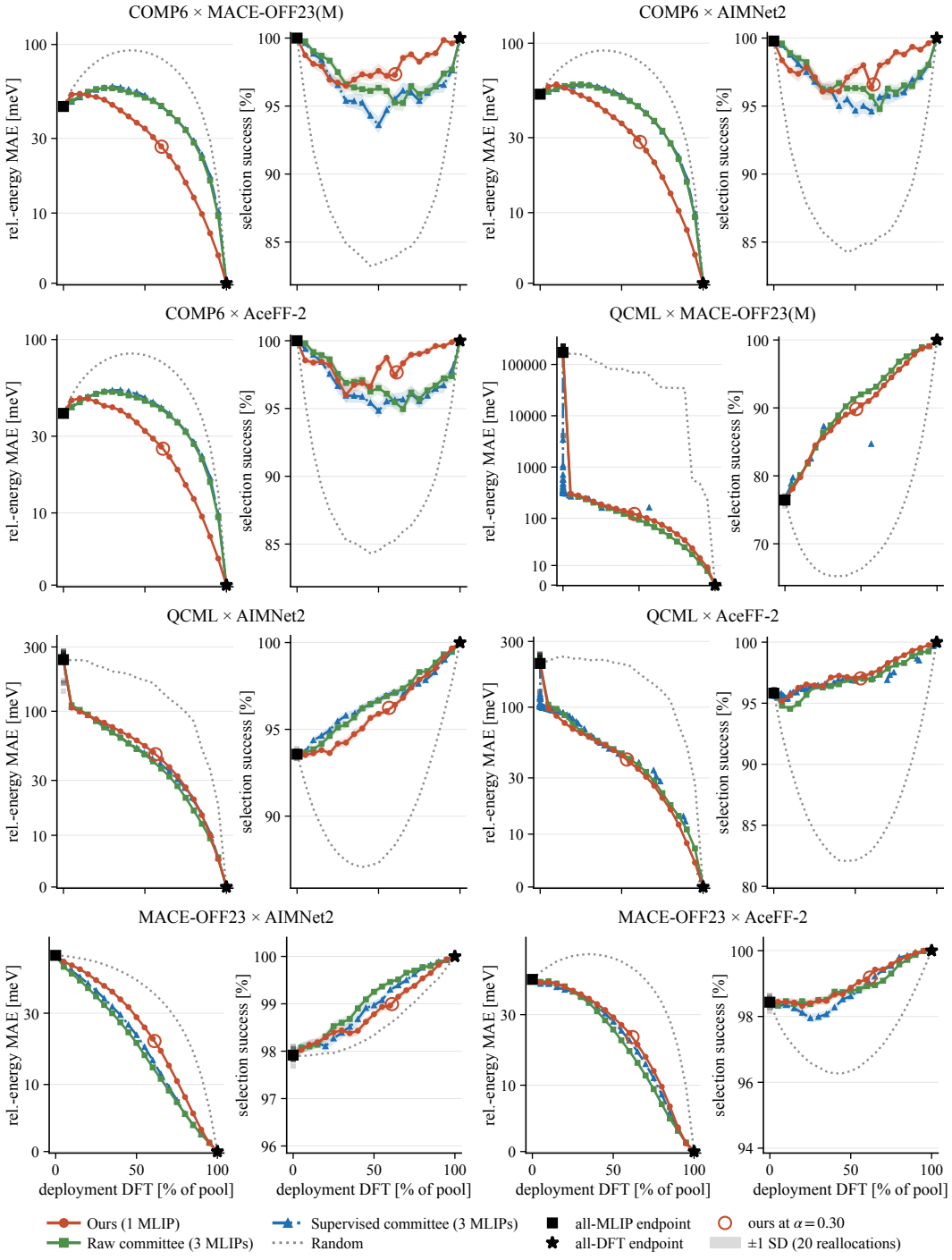

Figure 23: On COMP6, ours reaches a given relative-energy error and selection success with less deployment DFT than either committee; elsewhere the curves largely overlap. Per cell, relative-energy MAE (asinh axis, linear below 10 meV) and selection success against deployment DFT, with all-MLIP (square) and all-DFT (star) endpoints and a ring at ours, $\alpha = 0.30$. Means over 20 reallocations at 5% DFT steps, each split at its nearest attainable point; bands ± 1 SD. The saved sweep has no oracle. Success axes are zoomed per cell.

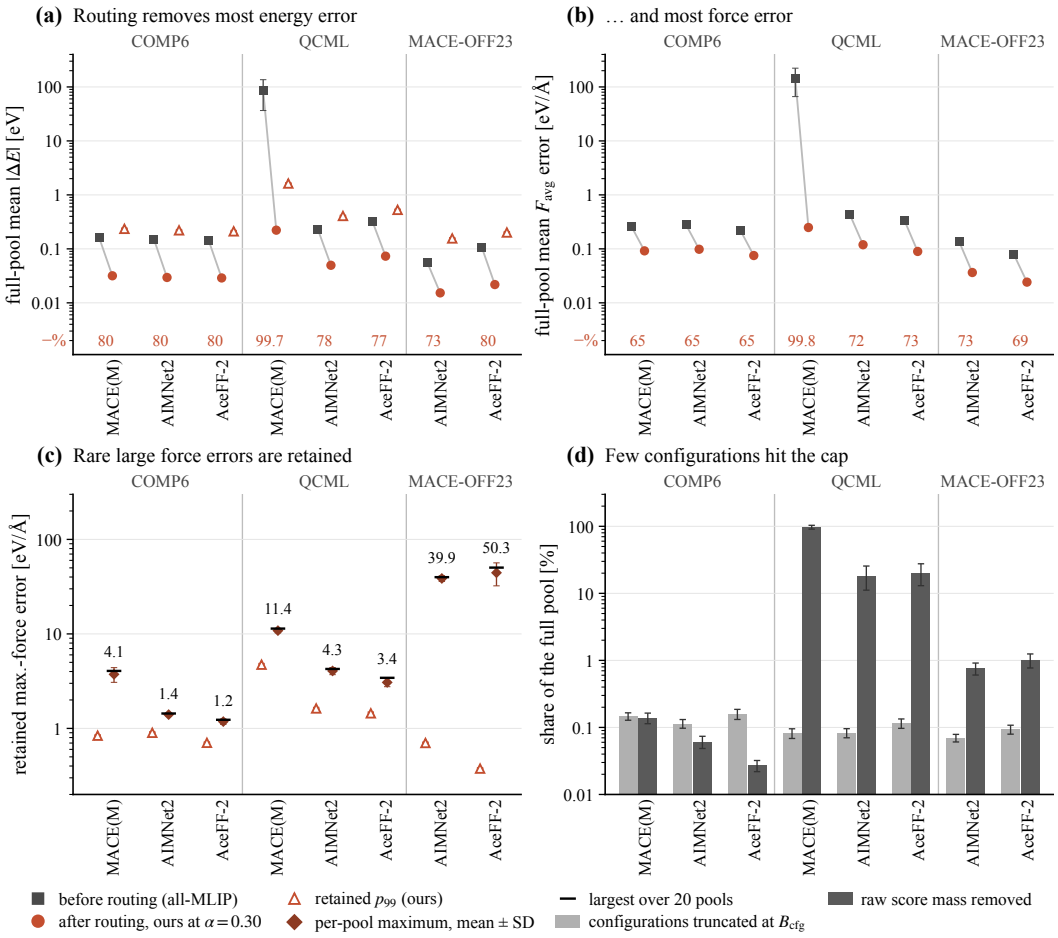

Figure 24: Routing at $\alpha = 0.30$ removes most physical error, but rare large force errors remain. (a) Full-pool mean aligned energy error and (b) mean atom-force error before routing and after routing with ours; numbers give the relative reduction (%), triangles the retained 99th percentile. (c) Retained maximum-force error: 99th percentile, per-pool maximum (mean ± SD), and largest over 20 pools. (d) Share of the pool truncated at B_{cfg}, as configurations and as raw score mass. Log axes. Uncapped tails are outside the guarantee.

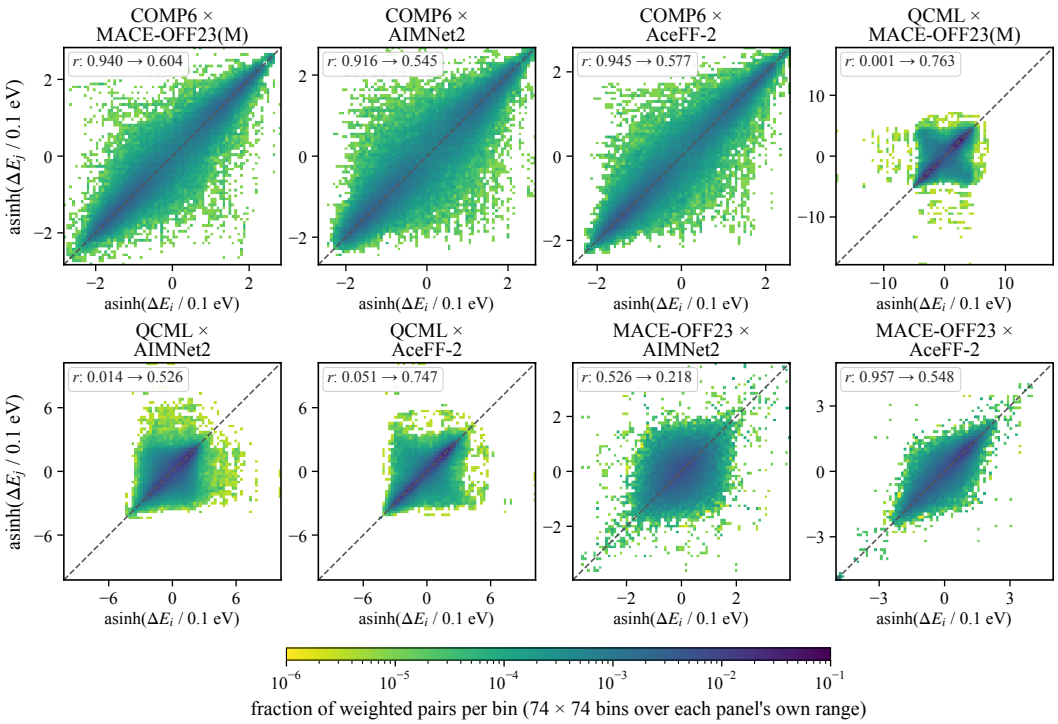

Figure 25: Signed residuals of one molecule's configurations are strongly correlated on COMP6 and MACE-OFF23 × AceFF-2, so energy differences benefit from cancellation that partial routing removes. Weighted density of signed aligned residual pairs $(\Delta E_i, \Delta E_j)$ within an identity (up to 64 pairs per eligible identity, both orientations), in $\mathrm{asinh}(\Delta E/0.1\,\mathrm{eV})$ coordinates; the dashed diagonal is exact cancellation. Annotations give the paired Pearson r before routing $\rightarrow$ after ours at $\alpha = 0.30$, averaged over 20 reallocations. One shared log color scale of pair fraction per bin.

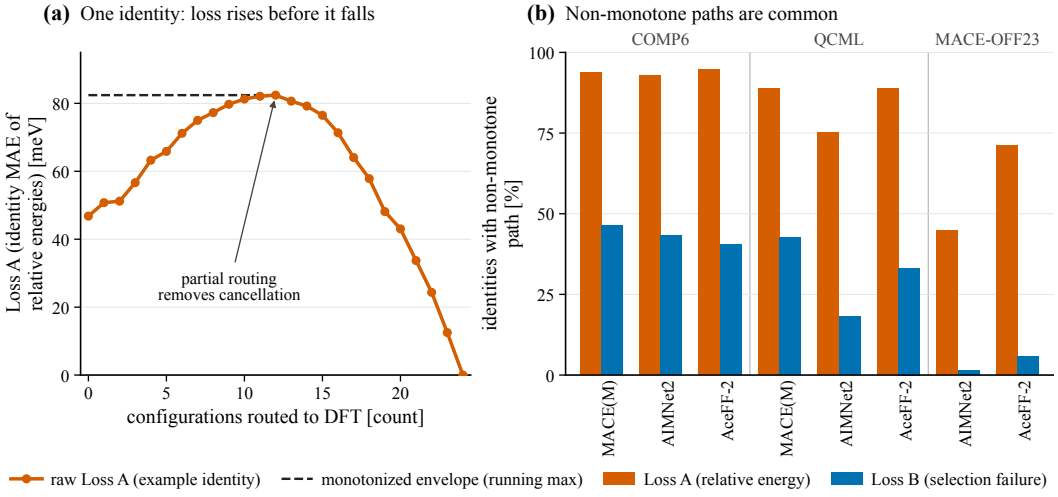

Figure 26: Partial routing can raise a molecule's relative-energy error before full DFT drives it to zero, and such paths are common. (a) One saved COMP6 × AceFF-2 identity: Loss A (identity MAE of relative energies) and its running-maximum envelope against the number of configurations routed; an illustration, not a prevalence estimate. (b) Per cell, the percentage of identities whose routing path is non-monotone for Loss A and for Loss B (selection failure).

K CERTIFYING OTHER LOSSES DIRECTLY

Remark 3 allows any bounded monotone loss in place of ℓ, so downstream objectives can be certified directly instead of through the proxy. We consider two identity-level losses on identities with at least two configurations. *Loss A* is the relative-energy MAE of the identity, capped at the largest fitting-identity path loss, with budget ϵ_A a multiple of each cell's fitting all-MLIP mean. *Loss B* is the indicator that the selected lowest-energy configuration lies more than 1 kcal/mol above the true minimum, with a probability budget ϵ_B. Because both can be non-monotone in the cutoff (Appendix J.1), we compare *M*, CRC on the running-maximum monotone envelope of the loss; *W-max* and *W-mean*, CRC with whole-identity routing by the maximum or mean configuration score; *L*, Learn-then-Test on the raw loss over a fixed fitting-set grid with $\delta = 0.10$; *P*, the original proxy rule at $\alpha = 0.30$, which carries no downstream certificate; and *N*, a naive empirical cutoff with no certificate. M and W bound the expectation, whereas L holds with probability at least 0.90 over the calibration draw, so their costs are not directly comparable. DFT percentages here count eligible non-singleton configurations.

Findings. Strict physical-unit losses, certified with the same score and CRC rule, cost almost all savings (Fig. 27): at the smallest budgets the equal-cell mean DFT is 55.0% for aligned energy (0.05 kcal/mol/atom), 85.3% for raw energy, and 94.3% for maximum force (0.005 eV/Å), and a strict binary loss needs 97.5–99.9%. The proxy guarantee does not transfer downstream: P exceeds the Loss-A budget (0.3× the fitting mean) in five of eight cells, with ratios 1.01–2.14, and the Loss-B budget (0.05) in one (Fig. 28c). Direct certification costs depend on the objective (Fig. 28a, b): at the main Loss-A budget, M, W-max, and L route 86.1%, 79.3%, and 96.7%, while at $\epsilon_B = 0.05$ they route 61.0%, 8.4%, and 21.8%; W-max is cheap partly because all-MLIP already meets the selection budget in several cells. Expectation-level methods can exceed the budget in individual repeats (Fig. 28d, e). Relative Loss-A budgets correspond to different absolute accuracies across cells. In particular, the saved QCML × MACE-OFF23(M) fitting mean is 14,761 eV, so its main budget is 4,428.5 eV, and on all three QCML Loss-A panels every certified method routes 100% (Fig. 29). We report these values as saved; the export does not establish the cause of the extreme scale.

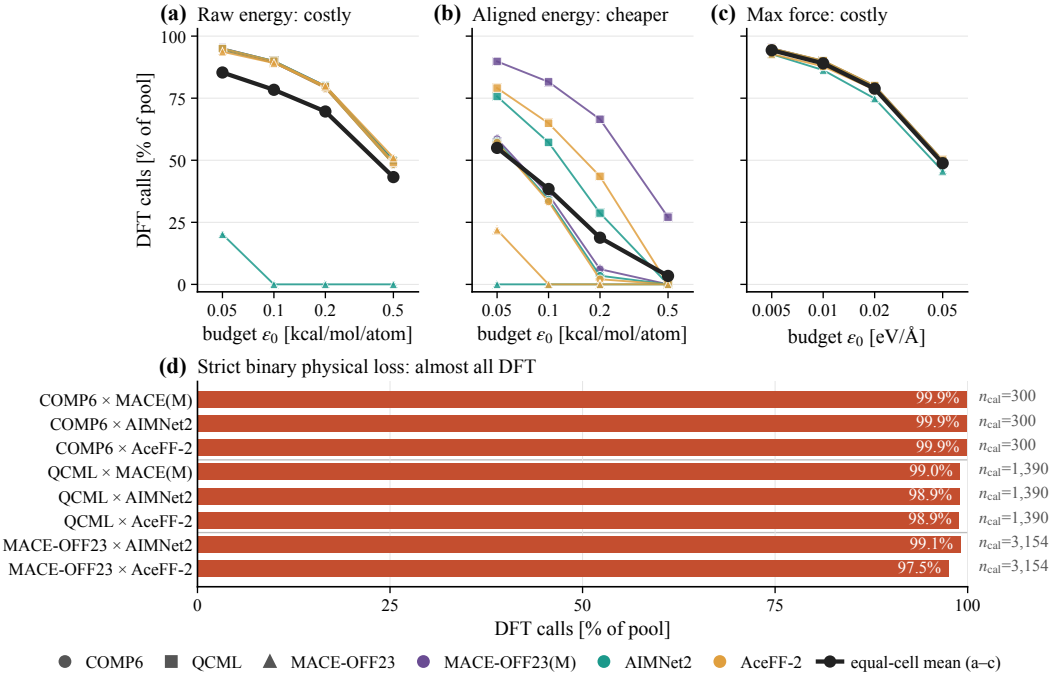

Figure 27: Strict physical-unit budgets cost almost all of the DFT savings. (a–c) DFT calls against an absolute budget for raw and aligned energy (kcal/mol/atom) and maximum force error (eV/Å), with the same structural score and CRC rule; thin lines are cells, the bold line the equal-cell mean of 20-reallocation means. (d) A strict binary loss (aligned $|\Delta E| > 1$ kcal/mol or maximum force > 0.10 eV/Å; budget 0.10 violations per identity) at each cell's calibration size. Thresholds are examples, not universal tolerances.

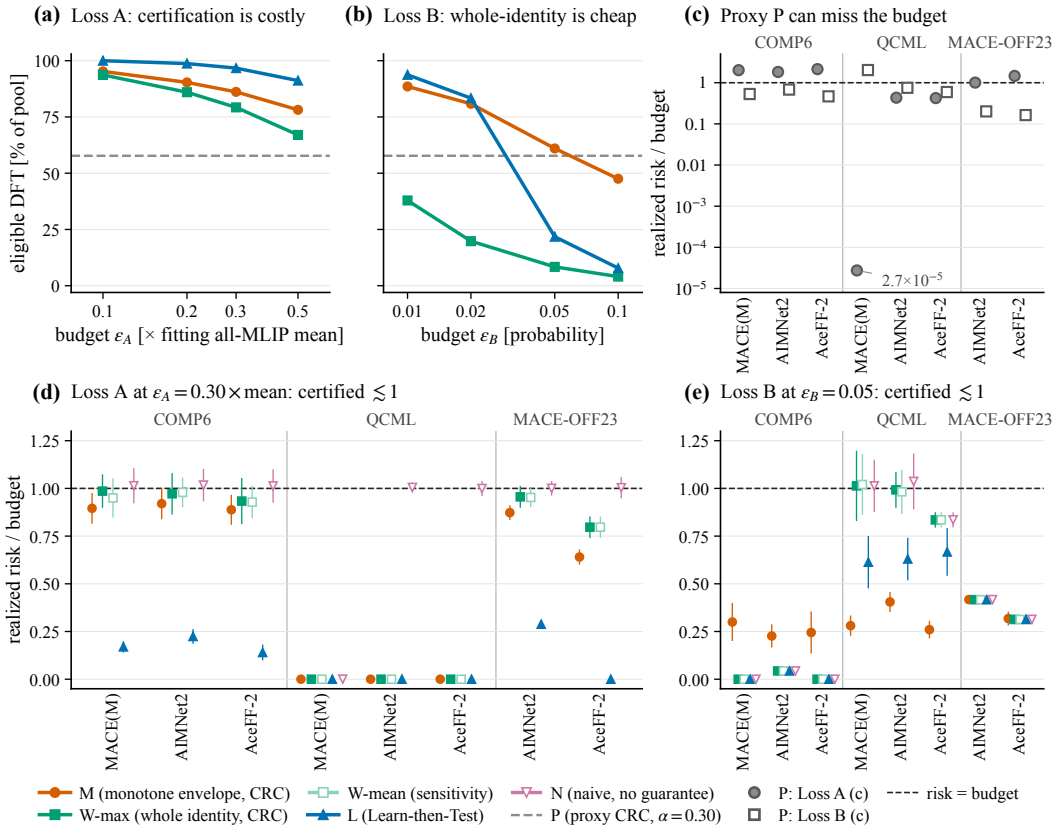

Figure 28: The proxy guarantee does not carry over to downstream losses, and direct control has its own cost. (a, b) Eligible DFT against the budget for Loss A and Loss B under M, W-max, and L; the dashed line is the uncertified proxy rule P at $\alpha = 0.30$. Equal-cell means of 20-repeat means. (c) Realized risk/budget of P per cell (log axis); the tiny QCML × MACE-OFF23(M) Loss-A ratio reflects its extreme fitting anchor. (d, e) Per-cell realized risk/budget of every method at the main budgets, mean ± SD over 20 repeats. L is high-probability; M and W are expectation-level.

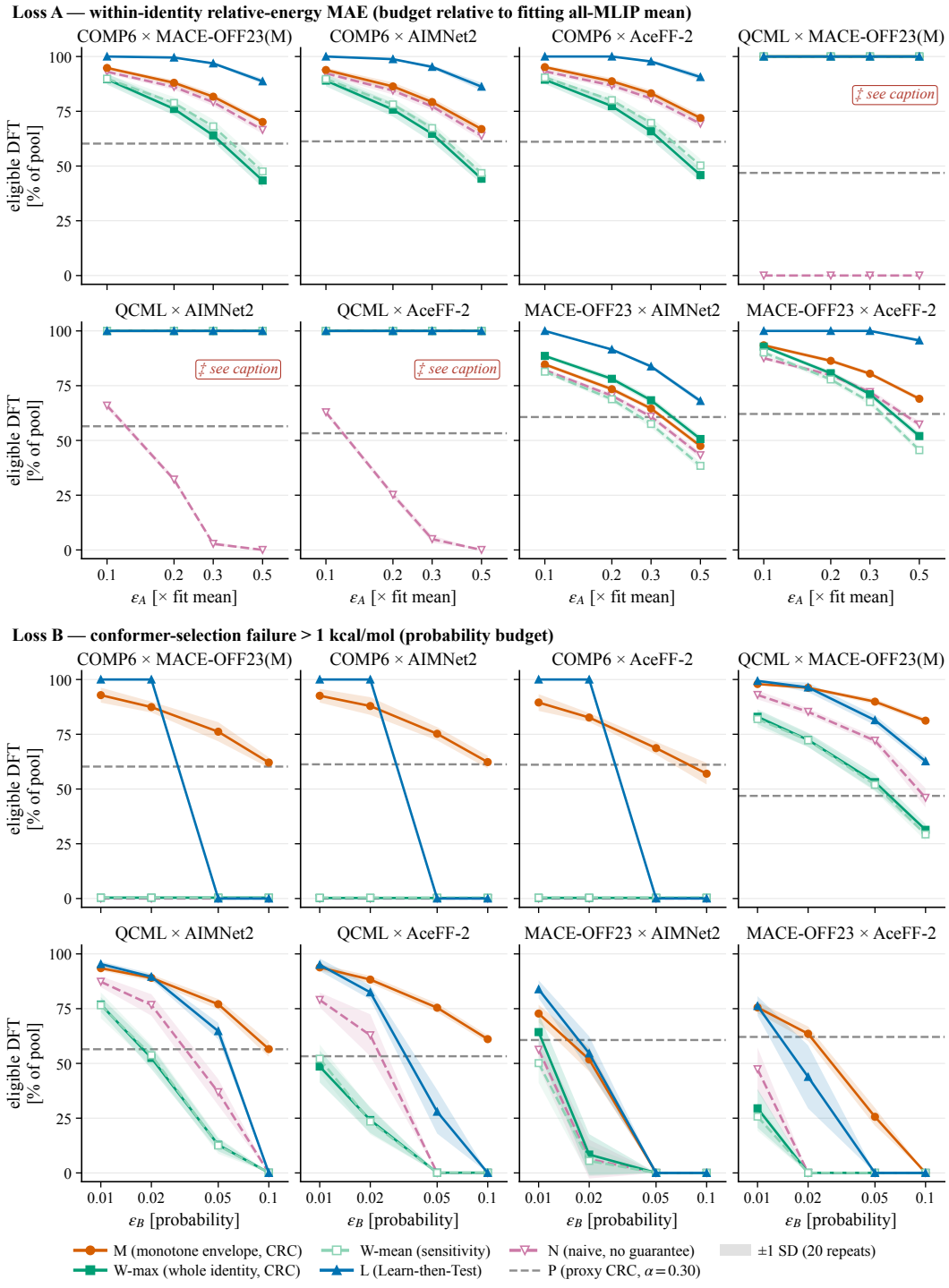

Figure 29: Downstream certification cost depends strongly on the cell and the objective. Eligible DFT against the budget in every cell, Loss A (top two rows) and Loss B (bottom two), for M, W-max, W-mean, L, and N (mean ±1 SD over 20 repeats); P is a horizontal reference. ‡ Data issue, plotted without alteration: the saved QCML × MACE-OFF23(M) Loss-A fitting mean is 14,761 eV (main budget 4,428.5 eV), and on all three QCML Loss-A panels every certified method routes 100%. Budgets are relative to fitting anchors, not a common absolute accuracy.

L LABEL COST AND TRANSFER

The reported savings are per reused pool and exclude the labels of $\mathcal{D}_{\text{fit}}$ and $\mathcal{D}_{\text{calib}}$. Charging them up front, and counting each cell in full-pool equivalents, the fitted selector breaks even after $P = 2$ comparable reused pools with 2.5% or 10% of identities for fitting and after $P = 4$ with 40% (Fig. 30); for a single pool it costs 1.17–2.07 full-pool equivalents, more than labeling the pool outright. Reuse assumes comparable pools, and the predictor must be fitted per population: a frozen structural score moved to another population keeps mean $G = 0.17$ (range -0.09 to 0.36, positive in 12 of 14 directed transfers) after target recalibration, against 0.80 in domain, and the QCML predictor anchored on COMP6 stays at $G \leq 0.15$ (Fig. 31).

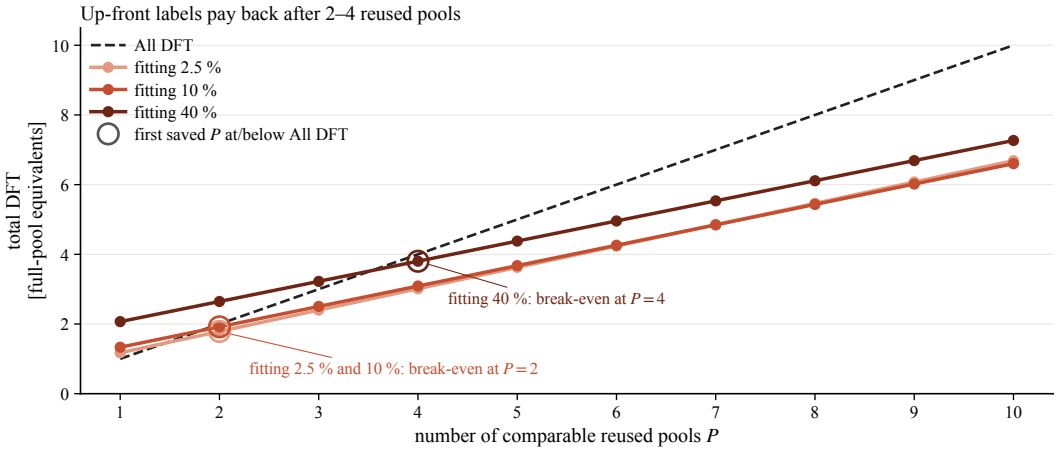

Figure 30: Up-front labels pay back after 2–4 reused pools. Total DFT in full-pool equivalents (fitting and calibration labels up front, plus routed deployment labels) against the number P of comparable reused pools, for fitting at 2.5%, 10%, and 40% of identities and for all-DFT ($y = P$). Equal-cell means, each cell normalized by one deployment pool; rings mark the first saved integer P at or below all-DFT. Cost-model projections in configuration counts, not measured wall time.

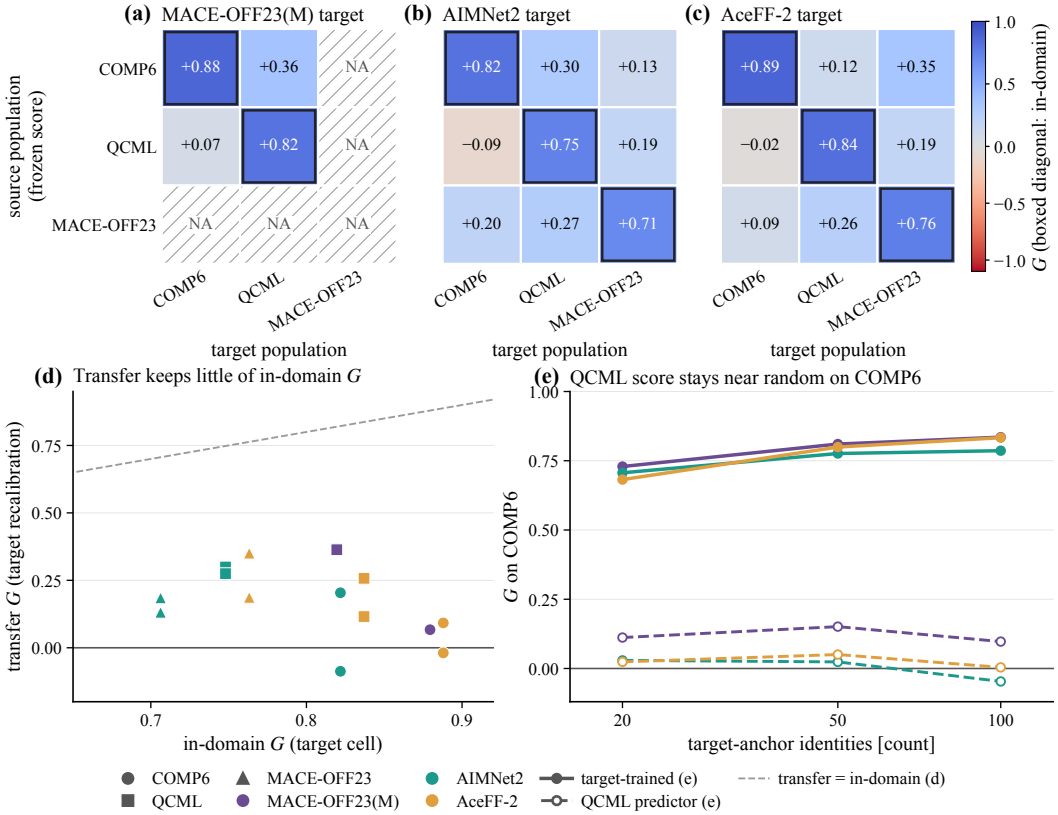

Figure 31: A frozen structural score transfers poorly across populations. (a–c) G of the frozen score from each source population (rows) on each target population (columns), per target MLIP, after target recalibration; boxed diagonals are in-domain, NA marks non-cells. (d) Transfer G against the target's in-domain G for all 14 directed transfers (dashed: equality). (e) Anchored QCML → COMP6: G against the number of target-anchor identities for the target-trained (solid) and frozen QCML (dashed) predictors. Target anchors and calibration labels are charged, so this is not zero-label deployment.

M LIMITATIONS

Scope of the guarantee. As is standard for conformal risk control, Proposition 1 is a marginal, expectation-level bound that holds within the dataset–MLIP cell used for fitting; a new population or target MLIP requires refitting and recalibration (Appendix L). The bound applies to the capped loss, and B_{cfg} can be set to an application tolerance. Where rare chemical families or single calibration draws must be protected, high-probability or group-conditional variants apply at additional DFT cost (Appendix K).

Cost accounting. Reported savings are per deployment pool. The one-time fitting and calibration labels are amortized after two to four comparable pools (Appendix L), so the method targets repeated screening of a population.

Reference mismatch. The residuals include differences between electronic-structure levels, which is the setting the method is designed for. When references match, committee disagreement is competitive, and the hybrid score combines both signals (Appendix G).

Downstream objectives. The proxy loss does not certify relative energies or conformer selection. Because the framework accepts any bounded monotone loss, these objectives can be certified directly (Appendix K).

Domain. The descriptor covers isolated molecules built from ten elements; periodic systems require extending its distance definitions.

N BROADER IMPACTS

Routing lowers the DFT cost of repeated molecular screening. It also replaces empirically chosen uncertainty thresholds with a stated, auditable loss and budget, which may help pretrained MLIPs enter workflows that require documented error control. Two risks follow. First, cheaper screening can enlarge campaigns, so total compute need not fall. Second, if routed labels are used to retrain the MLIP, regions the selector ranks low may remain unlabeled, and retraining requires refitting the predictor and recalibrating the threshold. The largest practical risk is overreading the guarantee as assurance of stable trajectories, correct conformer energies, or the absence of rare large errors, none of which it covers. Reports of routed results should state the controlled loss, its caps, and the calibration population.