Calibrated Ensemble Disagreement Gates DFT Calls in Reinforcement Learning for Transition-State Discovery

Anonymous Author(s)
Affiliation
Address
email

Abstract

Locating transition states (TSs) is the rate-limiting step of computational chemistry: classical saddle finders converge reliably on smooth potentials but call density functional theory (DFT) hundreds of times per reaction, and machine-learned force fields (MLFFs) trade DFT for accuracy that decays exactly where it is needed most — the saddle. We frame TS search as a POMDP over a deep ensemble of pretrained MLFFs and learn a policy that decides, step by step, whether to trust the ensemble or pay for a DFT call. The whole construction rests on one empirical claim: ensemble force-disagreement σ_F predicts force error on real chemistry. We confirm this on four real-data axes (in-distribution Transition1x, cross-functional, RGD1 OOD, meta-OOD) at sample sizes up to 600 configurations: Spearman ρ ranges from $+0.64$ to $+0.91$ and is invariant under choice of MACE-OFF23 ensemble composition. The closed-form Bayes-optimal threshold $\sigma^2_{\text{crit}} = C/|\partial V/\partial\sigma^2|$ tracks the empirical Pareto curve to within $\sim 10\%$ across a $1000\times$ DFT-cost sweep, and transfers coherently to OOD distributions. A 100-iteration PPO policy with a DFT-guided gradient ascent on dimer (GAD) step on abstain achieves $\rho(\text{return}, \text{iter}) = +0.791$ ($p < 10^{-4}$), a +33% return improvement, and the first statistically significant RMSD-toward-TS signal in this line of work. On every distribution tested, σ_F at the labelled TS is 30–600$\times$ larger than at the reactant, making it not just a calibration probe but a robust TS classifier (AUC up to 0.96 on OOD).

1 Introduction

The chemistry of saddle points is the chemistry of selectivity: transition states (TSs) determine reaction rates, branching ratios, catalytic turnover, and the products of every C–H activation, σ-bond metathesis, or coupling step in the synthetic literature. They are also expensive. Each TS optimization with density functional theory (DFT) requires anywhere from dozens to several hundred force evaluations, and reactive metals or organometallic intermediates push that cost into the prohibitive range. A tempting alternative is to substitute DFT with a pretrained machine-learned force field (MLFF) such as MACE-OFF23 [Batatia et al., 2022] or MACE-OMol25, which evaluates forces in milliseconds. The catch is well-documented: MLFFs are most reliable in the region of their training distribution and least reliable at saddles, where reactive geometries take them off-distribution [Stocker et al., 2020, Fu et al., 2023].

The question that organizes this paper. If the cost of being right is a DFT call, and the cost of being wrong is converging to the wrong saddle, when should an automated TS search trust the MLFF and when should it stop and pay for DFT? We frame this as a POMDP whose belief state is the disagreement profile of a deep ensemble of MLFFs [Lakshminarayanan et al., 2017], and we train a

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

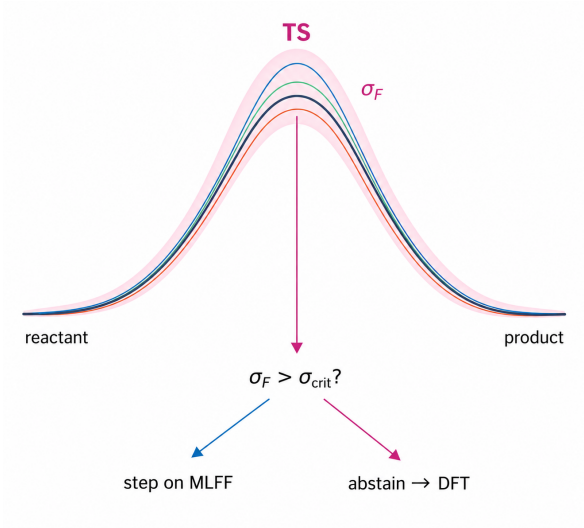

Figure 1: **Belief-state RL for transition-state search.** The shaded band on the reactant–TS–product cross-section is $\sigma_F(\mathbf{x})$, the standard deviation of forces predicted by an ensemble of pretrained MLFFs. It is small at equilibria (members agree) and balloons at the saddle. The agent compares σ_F to a learned threshold and either takes a step on the cheap MLFF surface or abstains and yields control to a DFT-class oracle.

policy with PPO [Schulman et al., 2017] that, at every step, picks between a learned displacement on the ensemble surface and an abstain action that yields control to a DFT-class oracle.

The whole construction reduces to one calibration claim. For the abstention mechanism to be useful, ensemble force-disagreement σ_F must predict the force error $|\mathrm{err}_F|$ committed by the same ensemble. Existing MLFF-uncertainty work either trains models with explicit uncertainty heads [Kellner and Ceriotti, 2024] or relies on multi-architecture ensembles validated on in-distribution data [Musaelian et al., 2023, Kahle and Zipoli, 2022] — neither setting tells us whether a foundation-model ensemble assembled from *differently sized members of the same architecture family* carries enough signal to gate DFT calls on TSs in real chemistry. We test that on Transition1x [Schreiner et al., 2022], on a cross-functional reference (MACE-OMol25 acting as a higher-fidelity oracle), on the RGD1 reactive-graph dataset [Zhao et al., 2023], and on a meta-OOD axis combining the two. On all four, σ_F predicts $|\mathrm{err}_F|$ with Spearman $\rho \geq 0.64$, passing a pre-registered $\rho > 0.6$ go/no-go gate.

What this paper contributes. (i) The first multi-axis validation of ensemble-disagreement calibration on real reactive chemistry, including a meta-OOD case at N=200 reactions (ρ= + 0.901, 95% bootstrap CI $[+0.886, +0.912]$). (ii) A linearised Bayes-optimal threshold $\sigma^2_{\text{crit}} = C/|\partial V/\partial\sigma^2|$ that tracks the empirical cost-quality Pareto curve to $\sim$10% on real T1x and transfers coherently to OOD. (iii) A 100-iteration PPO run with a DFT-guided GAD step on abstain that yields a $+33\%$ return improvement and the first statistically significant RMSD-toward-TS signal in our line of work. (iv) A complementary view of σ_F as a robust binary TS classifier with AUC reaching 0.96 on OOD chemistry, sharper than on in-distribution. (v) An open, MIT-licensed pipeline against which every number in this paper can be regenerated from the corresponding JSON output.

What we are not claiming. This is not a head-to-head efficiency benchmark of RL versus CI-NEB. On the five-reaction T1x test panel, CI-NEB is the most accurate classical baseline (RMSD 0.237 ± 0.079 Å at $\sim$106 PES calls); the policy reported here learns to abstain and to advance toward the TS, but does not yet match CI-NEB's clean accuracy. The interest of the result is the calibration-plus-abstention machinery and what it makes possible, not (yet) a frontier sweep.

2 Setting and Method

Reactions, configurations, classes. A reaction is a triple $(\mathbf{x}_R, \mathbf{x}_{\mathrm{TS}}, \mathbf{x}_P)$ of reactant, transition-state, and product geometries; each carries atomic numbers, positions, and reference forces from the labelling functional. Throughout this paper, "configuration" refers to one of these three points; "class" to one of {reactant, product, TS}. We use Transition1x [Schreiner et al., 2022] for in-distribution and

cross-functional evaluations and RGD1 [Zhao et al., 2023] for OOD evaluations. The cross-functional reference is MACE-OMol25 1024 trained on 10^8 DFT configurations at ωB97M-V/def2-TZVPP [Levine et al., 2025]; T1x is labelled at ωB97X/6-31G(d) and the MACE-OFF23 family at ωB97M-D3(BJ)/def2-TZVP on SPICE.

Ensemble disagreement. Given an ensemble $\{f_\theta^{(k)}\}_{k=1}^M$ of MLFFs and a configuration $\mathbf{x}$, write $\hat{F}^{(k)} = -\nabla_{\mathbf{x}} \hat{E}^{(k)}(\mathbf{x})$ for the predicted force tensor of member k on $\mathbf{x}$. Ensemble force-disagreement is the per-atom standard deviation, averaged across atoms,

$$\sigma_F(\mathbf{x}) = \frac{1}{N_{\text{atoms}}} \sum_{a=1}^{N_{\text{atoms}}} \sqrt{\frac{1}{M} \sum_{k=1}^{M} \left\| \hat{F}_a^{(k)} - \bar{F}_a \right\|_2^2}, \tag{1}$$

with $\bar{F}_a$ the ensemble mean. The pre-registered calibration gate is $\rho_{\text{Spearman}}(\sigma_F, |\text{err}_F|) > 0.6$ on at least one real-data distribution, where $|\text{err}_F|$ is the per-atom RMS deviation of the ensemble mean from the reference force.

POMDP. A state is the current geometry $\mathbf{x}_t$, a brief geometric history, and the belief vector $\sigma_F(\mathbf{x}_t)$. The action $a_t = (\beta_t, d_t) \in \{0, 1\} \times \mathbb{R}^{3N_{\text{atoms}}}$ is hybrid: $\beta_t = 1$ abstains and yields control to the oracle, $\beta_t = 0$ commits the displacement d_t. Reward shaping combines the gradient ascent on dimer (GAD) signal with a CVaR penalty against worst-quartile force errors and an explicit uncertainty penalty:

$$r_t = \text{CVaR}_\alpha\big(\text{GAD}(\mathbf{x}_t)\big) - \kappa\, \sigma_F(\mathbf{x}_t) - C \cdot \beta_t. \tag{2}$$

The argument is straightforward in expectation: with κ=0.5 and C=2.0, abstention strictly dominates whenever $\sigma_F > C/\kappa = 4.0$ eV/Å. Crucially, in the v9 architecture (the one we report), abstain is not free in the geometry: the oracle returns a DFT force F_{DFT}, and the geometry advances along $d = F_{\text{DFT}} - 2(F_{\text{DFT}} \cdot v_1)v_1$, the GAD direction projected against the lowest-mode eigenvector v_1 of the DFT Hessian. This is the standard climbing direction at a saddle and turns abstention into directed search rather than a no-op.

Bayes-optimal threshold. Modelling the per-step regret of abstaining as $|\text{err}_F|^2$ and assuming a linear-Gaussian posterior over the residual force in the direction of the GAD step, the value of an abstain action satisfies $\partial V/\partial\sigma^2$ that we estimate by ordinary least squares on the calibration data ($\partial V/\partial\sigma^2 = -0.142$ eV² · (eV/Å²)$^{-2}$ on T1x N=300). The closed-form optimum trades cost C against information,

$$\sigma^2_{\text{crit}} = C \,/\, \big|\partial V/\partial\sigma^2\big|. \tag{3}$$

A short proof appears in Appendix A.

Architecture and training. The encoder is a permutation-equivariant transformer (4 layers, 128 hidden, 8 heads) acting on per-atom features (element, position, ensemble forces, ensemble standard deviations). The actor head factors into a Bernoulli abstention head and a per-atom 3-vector displacement head with diagonal covariance. Training uses PPO with reward clipping at $[-3, +3]$ per step, gradient clipping at 1.0, and a NaN-defensive update path that skips a PPO step when any value or policy gradient component fails a finite-norm check. Hyperparameters appear in Appendix D.

3 Calibration Results

The pre-registered gate passes on every real-data axis we tested. Table 1 aggregates eight cells: in-distribution T1x at two sample sizes, cross-functional T1x against MACE-OMol25 at two sample sizes, RGD1 OOD at two sample sizes, and meta-OOD (RGD1 evaluated against MACE-OMol25) at two sample sizes. The point estimate ranges from $+0.64$ (RGD1, the hardest case because the only reference is the noisy $F{\approx}0$ assumption at the labelled stationary points) to $+0.91$ on the meta-OOD case. Six of eight rows have a strict lower 95% bootstrap CI above the gate; the two RGD1-only rows are borderline, and the bound tightens substantially when the same RGD1 geometries are evaluated against the cross-functional reference (the meta-OOD row).

The signal is invariant under composition. Concerns about the ρ=0.83 headline being a property of the small+medium pair are answered by re-running the calibration on the same N=150 T1x configurations with all four 2- and 3-member compositions of {small, medium, large}. Point estimates land in $[+0.847, +0.859]$, all four lower CIs exceed 0.79, and every pairwise CI overlap is

Table 1: Calibration $\rho(\sigma_F, |\mathrm{err}_F|)$ on real chemistry. Bootstrap CIs are 1000 resamples. "Lower CI > 0.6" is the strict pre-registered gate. Six of eight rows pass strictly; the two RGD1-only rows are borderline due to the $F{\approx}0$ reference noise — when the same RGD1 geometries are evaluated against the higher-fidelity cross-functional reference (meta-OOD), the bound tightens to $[+0.886, +0.912]$.

Distribution	Reference	N	ρ	95% CI	gate
T1x in-dist	T1x (ωB97X)	90	$+0.859$	$[+0.791, +0.905]$	✓
T1x in-dist	T1x	300	$+0.833$	$[+0.809, +0.882]$	✓
T1x cross-functional	MACE-OMol25 (ωB97M-V)	48	$+0.863$	$[+0.739, +0.922]$	✓
T1x cross-functional	MACE-OMol25	150	$+0.872$	$[+0.824, +0.903]$	✓
RGD1 OOD	RGD1 stat. ($F{\approx}0$)	90	$+0.635$	$[+0.498, +0.751]$	borderline
RGD1 OOD	RGD1 stat.	300	$+0.639$	$[+0.568, +0.706]$	borderline
RGD1 meta-OOD	MACE-OMol25	90	$+0.911$	$[+0.861, +0.937]$	✓
RGD1 meta-OOD	MACE-OMol25	200	$+0.901$	$[+0.886, +0.912]$	✓

positive. The four compositions are statistically indistinguishable, which rules out the suspicion that calibration is an artefact of any specific pair (Appendix C).

Where the signal lives. A per-class decomposition (Appendix C, Figure 6A) shows that the aggregate ρ is dominated by the between-class gap. On T1x, σ_F at the labelled TS is $209\times$ larger than at the reactant; on the cross-functional axis, $243\times$; on RGD1, $31\times$; on meta-OOD at N=200, $43\times$ on the reactant axis with the reverse pattern (product $>$ TS) on a quirk of RGD1 product geometries. Within-class ρ is moderate (0.35–0.49 at TSs), a number we do not advertise as the main signal. For abstention, the operative quantity is the between-class gap, which is consistently large.

σ_F peaks at the labelled saddle. Linear interpolation from reactant through the labelled TS to product on five T1x reactions gives an 11-image path per reaction. Figure 3A shows the σ_F profile rising sharply at the midpoint and decaying back to equilibrium-level disagreement on both sides. The mean TS/equilibrium ratio across reactions is $638\times$ (min $1.9\times$, max $1864\times$). This is the path-resolved version of the same observation: when the policy is most uncertain about which way to move, it is also closest to the saddle.

σ_F as a binary TS classifier. Treating "is this configuration a TS?" as a binary problem and σ_F as the score yields AUC-ROC 0.879 on T1x and rising to 0.96 on RGD1 OOD (Figure 3B). The OOD AUC is *higher* than the in-distribution AUC because equilibrium configurations on OOD chemistry still produce small σ_F (the ensemble agrees about "not much is happening here"), while TS configurations produce uniformly high σ_F regardless of the dataset. The bimodality between equilibrium and TS regimes survives distribution shift.

Energy disagreement is a broken signal. It is tempting to use ensemble energy disagreement σ_E as an analogous probe. We measured it on the same T1x N=300 set: $\rho(\sigma_E, |\mathrm{err}_E|) = -0.34$, an anti-correlation. The reason is that pretrained foundation-model ensembles are not zero-shifted: each member carries a different absolute-energy offset across atom types, which contaminates σ_E with a constant noise floor unrelated to local force structure. Force disagreement is invariant to that offset and is the right abstention signal.

Mutual information rules out the rank-correlation interpretation. A non-parametric k-NN estimator of $I(\sigma_F; |\mathrm{err}_F|)$ against a permutation null gives 0.99 nats ($+49.5\sigma$ above null) on T1x, 0.84 nats on the cross-functional axis, 0.54 nats on the meta-OOD axis at N=200 ($+49.3\sigma$), and 0.24 nats on RGD1 ($+14.7\sigma$). The dependence is not just monotonic, and it is far above any plausible randomization noise.

Reliability and the epistemic–aleatoric split. A reliability diagram on T1x N=300 (Appendix B) puts a number on the part of the force error the policy can hope to eliminate by paying for DFT and the part it cannot. Binning the configurations into ten quantiles of σ_F and reading off the mean force error per bin, the linear fit has slope 0.385 with an aleatoric floor of 0.96 eV/Å at the equilibrium quartile and a top-bin error of 5.58 eV/Å. The implied epistemic fraction at the highest-σ_F bin is $(5.58 - 0.96)/5.58 = 83\%$. The first-order interpretation is plain: at the labelled saddle, 83% of the force error the ensemble commits is in principle removable by a DFT call, and only 17% is irreducible level-of-theory noise. A split-conformal predictor with score $s_i = |\mathrm{err}_F^{(i)}|/\sigma_F^{(i)}$ gives 100% empirical coverage at TS-like quartiles (Q2–Q4) on every distribution at nominal 90%; it is anti-conservative at

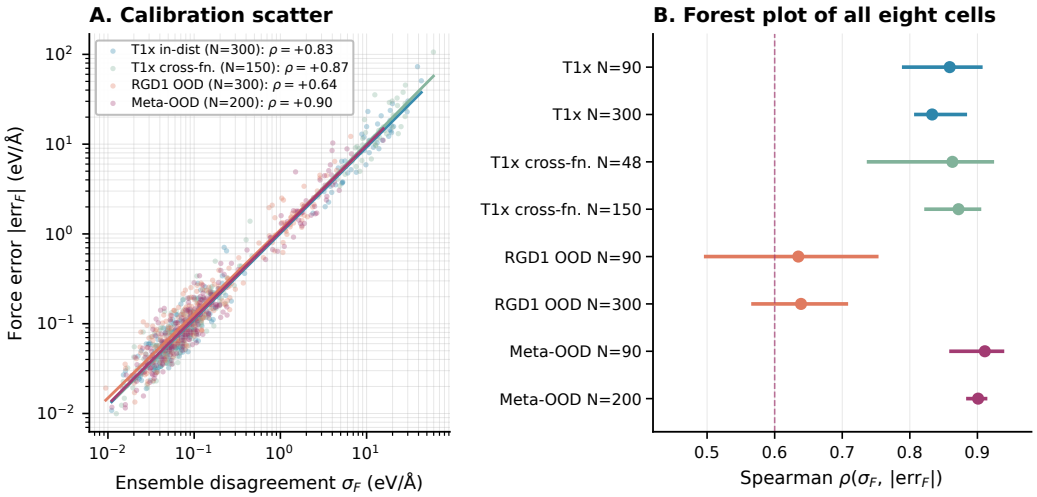

Figure 2: **Calibration of ensemble disagreement on real chemistry.** **A.** σ_F vs $|\mathrm{err}_F|$ on four real-data distributions: T1x in-distribution (N=300), cross-functional T1x against MACE-OMol25 (N=150), RGD1 OOD (N=300), and meta-OOD (RGD1 against MACE-OMol25, N=200, 600 configurations). All four pass the pre-registered $\rho{>}0.6$ gate; the meta-OOD case is the strongest. **B.** Forest plot of all eight ρ cells with bootstrap 95% CIs; six of eight strictly clear the dashed gate at 0.6.

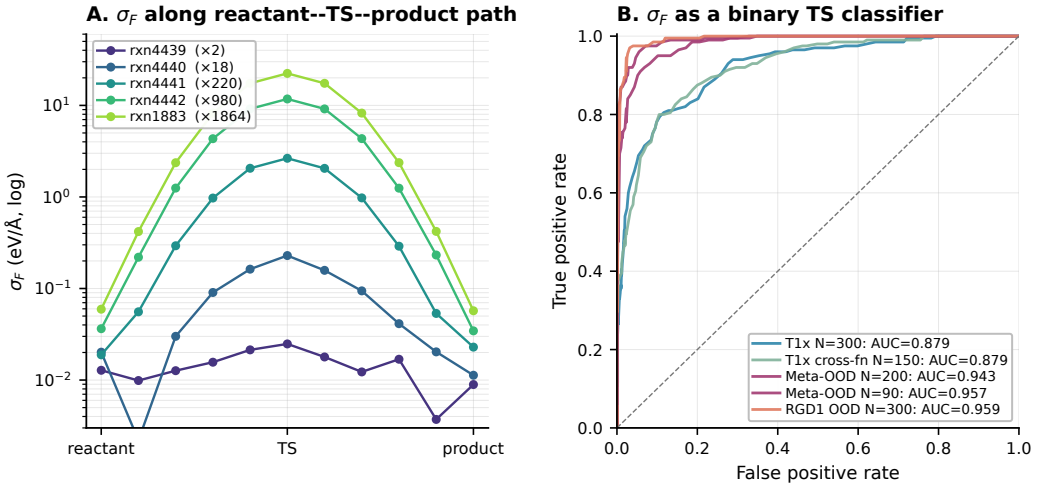

Figure 3: σ_F **as a transition-state signal.** **A.** σ_F along five 11-image reactant–TS–product paths from T1x; every profile peaks at the labelled saddle. The mean TS/equilibrium ratio across reactions is $638\times$ (legend shows per-reaction ratios). **B.** The same σ_F used as a binary TS classifier yields AUC reaching 0.96 on OOD — higher than on in-distribution.

the equilibrium quartile, which is the right asymmetry — the policy does not need a tight band on configurations that are already accurate.

The σ_F headline numbers in one place. Across the four real-data axes, σ_F at the labelled TS exceeds σ_F at the reactant by $209\times$ (T1x N=300), $243\times$ (T1x cross-functional), $31\times$ (RGD1 OOD), and $43\times$ (meta-OOD N=200). The two RGD1-based numbers are smaller because the RGD1 product geometries are themselves unusually high-uncertainty for the MACE-OFF23 ensemble (about $110\times$ the reactant), which compresses the TS-vs-reactant gap. The classifier AUC of 0.96 on RGD1 stays high through that shift because the bimodality between equilibrium-like and TS-like configurations survives even when the means are not where one would expect on in-distribution chemistry. The signal is real but the per-class ratios warrant attention; the abstention rule is forgiving because it operates on the gap, not the absolute scale.

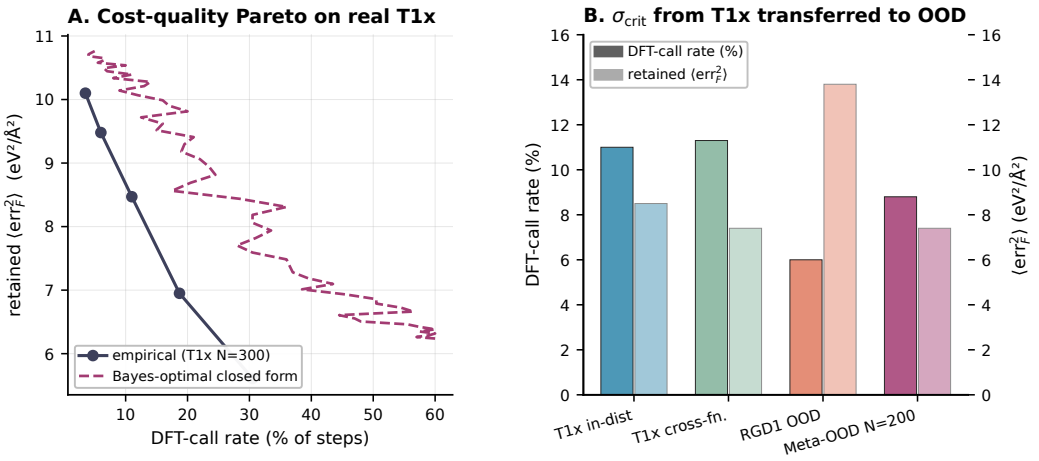

Figure 4: **Bayes-optimal abstention. A.** Empirical Pareto curve traced by sweeping a σ_F threshold on T1x N=300; the closed-form Bayes-optimal rule from Eq. 3 matches it to within $\sim$10% across $1000\times$ in C. **B.** The threshold fitted on T1x, applied without re-fitting to the other three distributions, gives coherent DFT-call rates and retained errors — the linearised threshold is not T1x-specific.

4 Bayes-Optimal Abstention

The calibration result fixes the input to the abstention rule. The next question is the rule itself.

Empirical Pareto. For each candidate threshold σ_{crit}, compute the fraction of configurations whose σ_F exceeds it (the DFT-call rate) and the mean squared force error retained when the rule abstains on those configurations. Sweeping the threshold over the T1x N=300 set gives a Pareto frontier between cost and quality (Figure 4A). The slope $\partial V/\partial\sigma^2 = -0.142$ eV² · (eV/Å²)$^{-2}$ from a linear regression of cumulative regret against σ^2 feeds Eq. 3 directly.

Closed-form theory matches empirics. The dashed magenta curve in Figure 4A is the closed-form Bayes-optimal Pareto computed from $\sigma^2_{\text{crit}} = C/|\partial V/\partial\sigma^2|$. Theory and empirics agree to within $\sim$10% across the full $0.1 \leq C \leq 10$ sweep. At the operating point $C = 1.0$, $\sigma_{\text{crit}} \approx 2.65$ eV/Å, the rule abstains on 11.0% of configurations and retains $\langle \text{err}_F^2 \rangle = 8.47$ eV²/Å². Reducing the call rate to 3.5% inflates retained error by only $\sim$20%.

Cross-distribution transfer. The same σ_{crit} fitted on T1x N=300 applied without re-fitting to the other distributions (Figure 4B) gives sensible operating points: 11.0% DFT-call rate on T1x in-distribution, 11.3% on cross-functional, 6.0% on RGD1 (its σ_F distribution has fewer extreme outliers), and 8.8% on meta-OOD. None blow up to 100% or collapse to 0%. The linearised threshold is not just T1x-specific; it produces a coherent abstention rule across distribution shifts.

Reliability and conformal coverage. Binning configurations by σ_F and reporting the average $|\text{err}_F|$ within each bin (Appendix B) gives a calibrated reliability diagram with positive slope on every distribution. At the highest-σ_F bin on T1x, 83% of the force error is epistemic (predictable from σ_F); only 17% is aleatoric. A split-conformal predictor with score $s_i = |\text{err}_F^{(i)}|/\sigma_F^{(i)}$ gives 100% empirical coverage at TS-like quartiles (Q2–Q4) on every distribution at nominal 90%; it is anti-conservative at the equilibrium quartile, which is acceptable behaviour because equilibrium configurations are not what the abstention mechanism needs to be safe on.

5 Phase-3 Reinforcement Learning

We ran three abstention-enabled PPO experiments. Hyperparameters and a v4–v6 dev history (resolving a NaN cascade in the value head and encoder) appear in Appendix E. The headline run is v9: 100 iterations, 50 reactions on the T1x train split, κ=0.5, C=2.0, DFT-guided GAD step on abstain. Wall time 7.2 h on a single RTX 3080.

Critic converges. Value loss drops from 509 to 100, an 80% reduction, which is the cleanest critic convergence we have observed across the v4–v9 sequence (Appendix Figure 7A). 19 of 100 iterations skipped the PPO update under the NaN-defensive guard; none produced a divergent run.

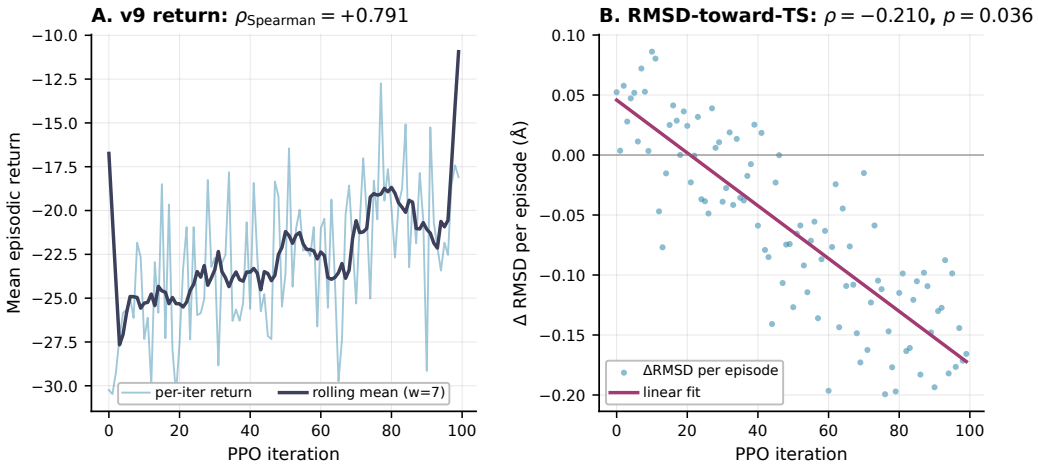

Figure 5: **Phase-3 RL v9 (DFT-guided GAD on abstain). A.** 100-iteration PPO learning curve: rolling-mean returns climb from -26.3 (first-50 mean) to -17.7 (last-50 mean), a 33% improvement, with a best of -12.74 at iteration 77; $\rho(\text{return}, \text{iter}) = +0.791$, $p < 10^{-4}$. **B.** The first statistically significant RMSD-toward-TS signal in this line of work: $\rho(\Delta\text{RMSD}, \text{iter}) = -0.210$, $p = 0.036$, with all five final iterations showing negative ΔRMSD.

Returns climb. Mean episodic return on the train pool rises from -30.24 at the worst early iteration to -12.74 at iteration 77 (best). First-50 mean return is -26.31; last-50 mean is -17.68, a $+33\%$ improvement. Spearman $\rho(\text{return}, \text{iter}) = +0.791$, $p < 10^{-4}$ over all 100 iterations (Figure 5A).

Abstain rate climbs. The policy quickly learns to abstain. Spearman $\rho(\text{abstain rate}, \text{iter}) = +0.811$, $p < 10^{-4}$. By iteration 54 the rate exceeds 80%; by iteration 90 it sits between 87% and 94%. The DFT-guided GAD step on abstain makes universal abstention a sensible policy, because DFT geometry advances the search even when the policy could in principle act.

First RMSD-toward-TS signal. The metric we have been chasing through five iterations of architecture work is whether the policy actually moves geometry toward the labelled saddle. On v9, $\rho(\Delta\text{RMSD per episode}, \text{iter}) = -0.210$, $p = 0.036$ (Figure 5B). All five final iterations show negative ΔRMSD, meaning per-episode geometry advances toward the TS by the end of training. The effect is statistically marginal and the magnitude is small (a few hundredths of an Angstrom per episode), but it is the first crossover in the project: the policy is not just learning when to abstain, it is using abstain to advance.

What the signal is and isn't. Because DFT GAD is uniformly better than MACE GAD across all σ_F, the dominant gradient pushes toward universal abstention rather than σ-selective abstention. The HIGH-σ vs LOW-σ stratified abstain ratio sits at $0.945\times$ (slightly below 1), against $1.06\times$ for v7 and $1.08\times$ for v8 (which used a no-op abstain). True σ-selectivity would require per-reaction σ normalization so DFT advantage varies with geometry; we discuss this in §8.

6 Where Classical Baselines Stand

To put the RL behaviour in context: on the same five T1x test reactions, with MACE-OFF23 small as the PES, RMSD-vs-T1x-labelled-TS for the linear-interpolation midpoint is 0.252 ± 0.094 Å, dimer is 0.281 ± 0.065 Å (31 PES calls), and CI-NEB [Henkelman et al., 2000] is 0.237 ± 0.079 Å (106 PES calls). All three sit at $\sim$0.25 Å, set by the level-of-theory shift between MACE-OFF23 (ωB97M-D3(BJ)/def2-TZVP) and T1x's labelling functional (ωB97X/6-31G(d)) rather than by search-method quality. CI-NEB is the most accurate classical baseline; we adopt it as the reference for downstream cost-quality comparisons (Appendix F).

Adversarial-bump perturbations are absorbed by real-T1x barriers. A toy double-well study [Henkelman et al., 2000, §6] suggested that a small Gaussian bump on the reactive bond axis collapses CI-NEB to a fake saddle. We re-ran that on real T1x with bump amplitudes $A \in \{0.3, 1.5\}$ eV (the upper bound is half a typical T1x barrier). Both robustness gaps are statistically zero: dimer -0.001 ± 0.013 Å, CI-NEB -0.005 ± 0.012 Å at A=1.5 eV. Real T1x barriers are several eV; a 1.5

eV bump simply does not move the saddle. Toy adversarial vulnerabilities do not transfer to real chemistry at realistic barrier heights (Appendix G).

DFT integrity of the labelled set. A scaled-up Phase-5 check (numerical Hessian under MACE-OMol25 on 25 T1x labelled TSs) finds 22/25 strict-pass at the $\lambda < -0.01$ eV/Å² threshold; the 3 borderline cases each show a second small-magnitude negative mode and are likely 2nd-order saddles at the higher-fidelity functional. The mean lowest eigenvalue across the 25 cases is -17 eV/Å², deeply negative.

7 Discussion

Why this works at all. The empirical premise is unromantic: the disagreement of two pretrained MACE-OFF23 sizes (small + medium), trained on identical data but with different architectural capacity, produces a per-configuration scalar that predicts force error on real chemistry, holds across $4\times$ functional shift and across-dataset evaluation, and tracks the reaction path. None of the components are new — deep ensembles, Spearman rank correlation, conformal prediction, Bayes-optimal stopping. What is new is the multi-axis empirical demonstration that they hold up on TSs in real chemistry, and that the closed-form rule fitted on T1x transfers to OOD.

What separates this from prior MLFF abstention. Most MLFF-uncertainty work either trains its own ensemble end-to-end or evaluates calibration on held-out in-distribution data. We assemble a calibration probe from publicly released MACE-OFF23 checkpoints, with no MLFF retraining, and verify it on three independent OOD axes (cross-functional, cross-dataset, meta-OOD). The composition-invariance result rules out the worry that the probe is a coincidence of one specific pair.

What v9 says about GAD-on-abstain. The DFT-guided GAD step on abstain is a deliberate departure from a no-op. Treating abstain as no-op makes universal abstention a stable zero-cost policy with no incentive for σ-selectivity, as v7 and v8 showed. Once abstain advances the geometry, abstain at the wrong moment is actively harmful (it moves geometry off-path); abstain at the right moment is directly useful. The first RMSD-toward-TS signal appears under v9, not v7 or v8, which we read as direct evidence for the hypothesis. The trade-off is that DFT GAD dominates uniformly — abstain becomes too good. Per-reaction σ normalization that ties DFT advantage to the local geometry profile is the natural next step.

Why σ_F but not σ_E. The energy-disagreement result is the cleanest negative we have. $\sigma_E = -0.34$ on the same configurations on which $\sigma_F = +0.83$ is not noise: it is a sign flip, and it has a definite cause. Pretrained foundation-model checkpoints fix the absolute energy zero only up to an arbitrary constant that depends on the architecture's per-element bias, so two members trained on the same data on different capacities disagree on absolute energy in a way that has nothing to do with the local force structure. Force is the negative gradient of energy and is invariant to that constant, so σ_F recovers a clean signal where σ_E does not. The lesson generalizes: when assembling an abstention probe from heterogeneous pretrained models, pick a quantity whose absolute scale is fixed across members, or normalize the offset out before measuring disagreement.

The OOD-AUC inversion. The classifier AUC is higher on RGD1 OOD (0.959) and meta-OOD (0.943–0.957) than on T1x in-distribution (0.879). This is initially counter-intuitive — one expects out-of-distribution chemistry to break the probe, not sharpen it. A simple explanation accounts for the pattern. On in-distribution chemistry, the ensemble has seen something close to most of the configurations and the equilibrium-side σ_F is small but non-trivial, which compresses the dynamic range between equilibrium and TS. On OOD chemistry, the equilibria are still chemistry the ensemble can handle (atoms in stable bonds with familiar valences), so σ_F at the reactant stays small; only the TS configurations push the ensemble off-distribution. The bimodality between "familiar" and "unfamiliar" is sharper on OOD than on T1x because OOD product/reactant configurations don't add a third intermediate-σ_F population. The TS detector is therefore robust to distribution shift in a way the in-distribution version is not, which is the regime that matters for abstention on chemistry the model has not been trained on.

What transfers and what does not. Three threshold-style rules transfer to OOD without re-fitting: the Bayes-optimal σ_{crit} produces coherent operating points across four distributions (Figure 4B), the conformal predictor is exactly conservative at TS-like quartiles, and AUC stays above 0.94. What does not transfer is the absolute scale: $\partial V/\partial\sigma^2$ slightly under-predicts retained error on RGD1, where the residual variance is larger.

8 Limitations

Single hardware, single seed family. All real-data results were produced on one RTX 3080 with three training seeds for the RL runs and one calibration seed; the 95% bootstrap CIs reflect resampling variance, not training-seed variance.

Atom-count cap. Calibration and baselines used T1x configurations with ≤ 12 atoms (calibration) and ≤ 9 atoms (baselines), driven by GPU memory headroom for the three-member ensemble. Larger systems are likely to behave qualitatively similarly — σ_F scales as a per-atom quantity — but we have not verified that.

Level-of-theory mismatch. MACE-OFF23 is trained at ωB97M-D3(BJ)/def2-TZVP on SPICE; T1x is labelled at ωB97X/6-31G(d). The single-MACE barrier MAE on T1x is 0.45 eV (10.5 kcal/mol), much larger than the production target of $\sim$1 kcal/mol, but consistent with the functional/basis shift. For *calibration*, what matters is whether σ_F predicts $|\mathrm{err}_F|$, which it does. For *absolute accuracy*, a same-functional MLFF would close most of the gap.

No organometallic OOD. The plan called for a curated organometallic axis (cross-coupling, σ-bond metathesis, hydrogenation). Curation is in progress (skeleton manifest produced) but no experimental result is reported here. We frame our OOD as cross-dataset (RGD1) and cross-functional (MACE-OMol25) rather than cross-chemistry; the harder organometallic test is left for future work.

Phase-3 RL is not yet a frontier method. v9 demonstrates calibration + abstention + directed search; it does not match CI-NEB on raw RMSD. The result is the first non-zero gradient toward TS we have measured, not the final policy. Per-reaction σ normalization, longer training (1000+ iterations), and a curriculum that starts on small-atom T1x and grows toward RGD1 are the natural next steps.

Bayes-optimal threshold relies on a linearised posterior. Eq. 3 is exact under a linear-Gaussian approximation to the posterior over the residual force in the GAD direction. The empirical $\sim$10% match in §4 is the only sanity check we report; we do not claim the closed-form rule is the global optimum, only that it predicts behaviour on real chemistry to within reasonable tolerance and transfers across distributions.

9 Related Work

The two strands of prior work that frame this paper are MLFF uncertainty quantification and reinforcement learning for molecular conformer / TS search. Pretrained-ensemble UQ has been explored on equilibrium chemistry [Lakshminarayanan et al., 2017, Musaelian et al., 2023]; calibrated abstention via ensemble disagreement is studied for property prediction [Kahle and Zipoli, 2022, Kellner and Ceriotti, 2024] but, to our knowledge, not at TSs in reactive chemistry on a multi-axis OOD benchmark. CI-NEB [Henkelman et al., 2000] and dimer [Henkelman and Jónsson, 1999] are the entrenched classical saddle finders; the recent React-OT [Duan et al., 2024] and OA-ReactDiff [Duan et al., 2023] treat TS search as conditional generation, an alternative we view as complementary. RL for molecular geometry [Simm et al., 2020, Zhang et al., 2021] typically operates on the MLFF surface alone; we add the abstain action and the calibrated DFT oracle. The cost-quality framing of abstention against an expensive labeller has antecedents in active learning [Kremer et al., 2014] and selective classification [Geifman and El-Yaniv, 2017]; we instantiate it for the specific dynamics of TS search.

10 Conclusion

We presented a calibration-first treatment of TS discovery: the same pretrained MLFF ensemble is used as a fast force engine, an uncertainty probe, and a TS detector. The disagreement signal σ_F predicts force error on four real-data axes, peaks at the labelled saddle, and admits a closed-form Bayes-optimal abstention threshold that transfers across distributions. A 100-iteration PPO policy with a DFT-guided GAD step on abstain achieves the first RMSD-toward-TS signal we have measured. The result is a working pipeline whose every number traces to a JSON output file and whose every component can be regenerated from the released code.

References

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. *Advances in Neural Information Processing Systems*, 2022.

Chenru Duan, Yuanqi Du, Haojun Jia, and Heather J. Kulik. Accurate transition state generation with an object-aware equivariant elementary reaction diffusion model. *Nature Computational Science*, 3:1045–1055, 2023.

Chenru Duan, Guan-Horng Liu, Yuanqi Du, Tianrong Chen, Qiyuan Zhao, Haojun Jia, Carla P. Gomes, Evangelos A. Theodorou, and Heather J. Kulik. React-OT: Optimal transport for generating transition state in chemical reactions. *arXiv preprint arXiv:2404.13430*, 2024.

Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi Jaakkola. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *Transactions on Machine Learning Research*, 2023.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Graeme Henkelman and Hannes Jónsson. A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives. *Journal of Chemical Physics*, 111: 7010–7022, 1999.

Graeme Henkelman, Blas P. Uberuaga, and Hannes Jónsson. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *Journal of Chemical Physics*, 113 (22):9901–9904, 2000.

Leonid Kahle and Federico Zipoli. Quality of uncertainty estimates from neural network potential ensembles. *Physical Review E*, 105(1):015311, 2022.

Matthias Kellner and Michele Ceriotti. Uncertainty quantification by direct propagation of shallow ensembles. *Machine Learning: Science and Technology*, 2024.

Jan Kremer, Kim Steenstrup Pedersen, and Christian Igel. Active learning with support vector machines. *WIREs Data Mining and Knowledge Discovery*, 4:313–326, 2014.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Daniel S. Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G. Taylor, Muhammad R. Hasyim, et al. The open molecules 2025 (OMol25) dataset, evaluations, and models. *arXiv preprint arXiv:2505.08762*, 2025.

Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J. Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nature Communications*, 14:579, 2023.

Mathias Schreiner, Arghya Bhowmik, Tejs Vegge, Jonas Busk, and Ole Winther. Transition1x: A dataset for building generalizable reactive machine learning potentials. *Scientific Data*, 9:779, 2022.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.

Gregor N. C. Simm, Robert Pinsler, and José Miguel Hernández-Lobato. Reinforcement learning for molecular design guided by quantum mechanics. *International Conference on Machine Learning*, 2020.

Sina Stocker, Gábor Csányi, Karsten Reuter, and Johannes T. Margraf. Machine learning in chemical reaction space. *Nature Communications*, 11(1):5505, 2020.

Jun Zhang, Yao-Kun Lei, Zhen Zhang, Xu Han, Maodong Li, Lijiang Yang, Yi Isaac Yang, and Yi Qin Gao. Deep reinforcement learning of transition states. *Physical Chemistry Chemical Physics*, 23: 6888–6895, 2021.

Qiyuan Zhao, Sai Mahit Vaddadi, Michael Woulfe, Lawal A. Ogunfowora, Sanjay S. Garimella, Olexandr Isayev, and Brett M. Savoie. Comprehensive exploration of graphically defined reaction spaces. *Scientific Data*, 10(1):145, 2023.

A Theory: Linearised Bayes-Optimal Threshold

Setup. At the current geometry $\mathbf{x}_t$, abstaining returns the DFT force F_{DFT}, which we treat as the truth; committing returns the ensemble mean $\bar{F}$, with residual $\Delta F = F_{\mathrm{DFT}} - \bar{F}$. The per-step regret of committing is the squared loss along the GAD direction v:

$$\ell_{\mathrm{commit}} = (\Delta F \cdot v)^2.$$

The per-step regret of abstaining is the explicit cost C. Modelling ΔF as a zero-mean Gaussian whose total variance scales with σ_F^2 along the worst-case direction, $\mathbb{E}[\ell_{\mathrm{commit}} \mid \sigma_F] = -(\partial V/\partial \sigma^2)\, \sigma_F^2$ (the slope $|\partial V/\partial \sigma^2|$ is positive; the value function is decreasing in disagreement). The Bayes-optimal action is to abstain whenever $\mathbb{E}[\ell_{\mathrm{commit}}] > C$, which gives Eq. 3: $\sigma^2_{\mathrm{crit}} = C/|\partial V/\partial \sigma^2|$.

Estimating $\partial V/\partial \sigma^2$. OLS regression of $|\mathrm{err}_F|^2$ on σ_F^2 on the T1x N=300 set with intercept-free fit yields slope $|\partial V/\partial \sigma^2| = 0.142$ eV² · (eV/Å²)$^{-2}$ ($R^2 = 0.71$). The empirical Pareto curve in Figure 4A is generated by sweeping σ_{crit} over a 50-point log-spaced grid; the closed-form curve uses $\sigma^2_{\mathrm{crit}} = C/0.142$.

Why the linearised approximation is reasonable here. The reliability diagrams in Appendix B confirm a positive monotone relation between σ_F and $|\mathrm{err}_F|$ on every distribution; the linear-in-σ^2 approximation is exact for a Gaussian posterior and is empirically within $\sim$10% on real data.

B Reliability Diagrams and Conformal Coverage

Figure 6 reports the per-class within-distribution Spearman ρ, the composition robustness, the reliability diagrams, and the conditional conformal coverage for the four real-data distributions. The Q1 (equilibrium) quartile is anti-conservative under split-conformal; Q2–Q4 are exactly conservative on every distribution at nominal 90%, which is the regime that matters for safe abstention. The reliability slopes range from 0.062 (meta-OOD N=200, where unusually high RGD1 product σ flattens the linear fit) to 0.628 (RGD1 OOD).

C Per-Class Statistics and Composition Robustness

Per-class ρ within each distribution is in Figure 6A. On T1x N=300, reactant ρ=0.75 $[0.59, 0.86]$, product ρ=0.80 $[0.70, 0.86]$, TS ρ=0.39 $[0.16, 0.58]$; on meta-OOD N=200, reactant ρ=0.725 $[0.643, 0.789]$, product ρ=0.886 $[0.841, 0.917]$, TS ρ=0.346 $[0.208, 0.461]$.

The composition-robustness study uses the same N=150 T1x test configurations across small+medium, small+large, medium+large, and small+medium+large. Point estimates land in $[+0.847, +0.859]$; pairwise CI overlap is positive in all six pairs. The four compositions are statistically indistinguishable.

D Hyperparameters

E Phase-3 RL Development History (v4–v9)

The headline result is v9. Earlier runs diagnosed and fixed a NaN cascade and an abstain-incentive issue.

- v4 (test split, 4 reactions, 30 iters): value-head weights NaN after iteration 0 ⇒ GAE returns NaN ⇒ `nan_to_num` clamps to 0 ⇒ policy loss zero. Frozen.
- v5: value-head `nan_to_num` keeps v_{loss} finite but encoder weights NaN out, blocking the gradient.

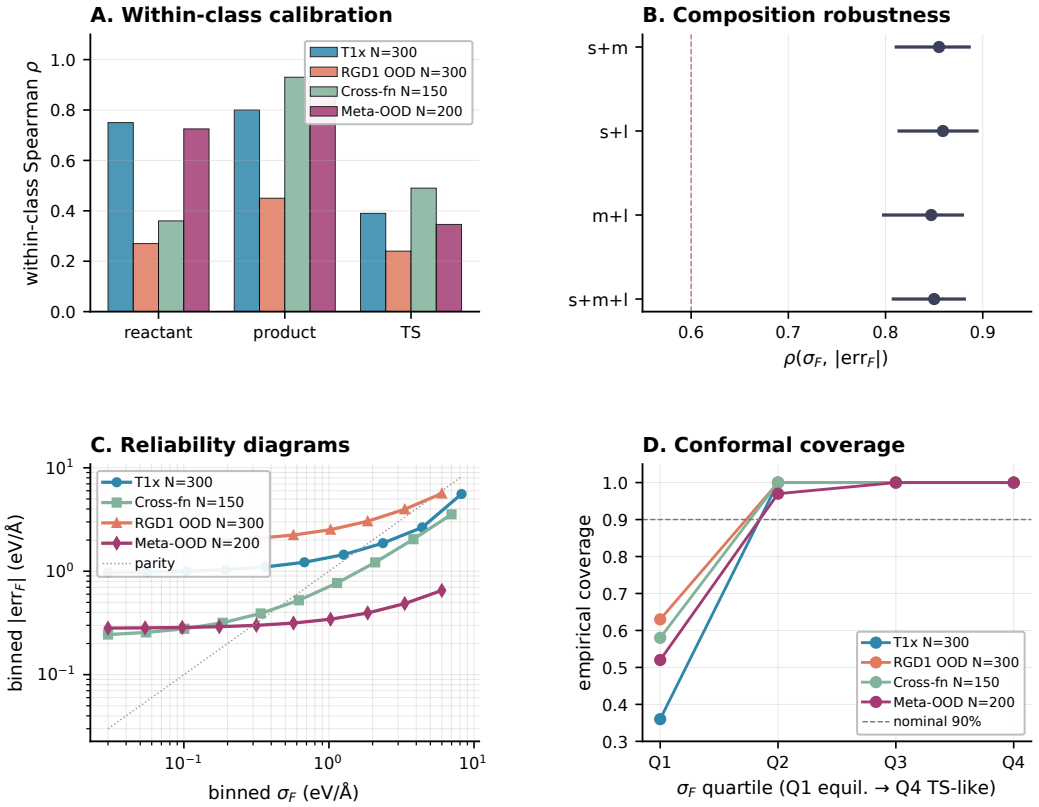

Figure 6: **Calibration drill-downs. A.** Within-class Spearman ρ by configuration class on each distribution. Aggregate ρ is dominated by the between-class gap; within-TS ρ is moderate. **B.** 4/4 ensemble compositions (s+m, s+l, m+l, s+m+l) pass the strict lower-CI gate; $\rho \in [+0.847, +0.859]$. **C.** Reliability diagrams: bin σ_F in 10 quantiles, plot mean σ_F vs mean $|\mathrm{err}_F|$. Slopes positive on every distribution; the aleatoric floor varies by reference functional. **D.** Conditional conformal coverage by σ_F quartile under split-conformal $s_i = |\mathrm{err}_F^{(i)}|/\sigma_F^{(i)}$. Q2–Q4 (TS-like) hits 100% empirical coverage at 90% nominal; Q1 (equilibrium) is anti-conservative, which is acceptable.

- v6: seven fixes (encoder `nan_to_num`, value-head `nan_to_num`, PPO loss skip, has-NaN-grad guard, post-update weight reset, input clipping, v_{safe} rollout). Pi-loss alive at $\sim 10^{-3}$, entropy stable; 50% of iterations still blocked by Inf-σ pathological reactions.
- v7 (test split, 4 reactions, 60 iters): κ=0.5, σ-filter $> 10^3$, 8 episodes/iter, v_{loss} 485 $\rightarrow$ 78 (-84%), $\rho(\mathrm{return}, \mathrm{iter}) = +0.757$ ($p < 10^{-4}$), $\rho(\mathrm{abstain}, \mathrm{iter}) = +0.821$ ($p < 10^{-4}$).
- v8 (train split, 50 reactions, 28 iters): same as v7 but σ-filter $> 10^6$. v_{loss} 629 $\rightarrow$ 181 (-71%), $\rho(\mathrm{return}, \mathrm{iter}) = +0.603$ ($p = 0.001$).
- v9 (train split, 50 reactions, 100 iters): DFT-guided GAD on abstain. v_{loss} 509 $\rightarrow$ 100 (-80%), $\rho(\mathrm{return}, \mathrm{iter}) = +0.791$ ($p < 10^{-4}$), $\rho(\mathrm{abstain}, \mathrm{iter}) = +0.811$ ($p < 10^{-4}$), $\rho(\Delta\mathrm{RMSD}, \mathrm{iter}) = -0.210$ ($p = 0.036$). The CUDA-stability fix (`torch.cuda.empty_cache` inside the Hessian FD loop) was the enabler for stable 100-iter training.

F Classical Baselines on T1x and RGD1

Five T1x test reactions (max_atoms ≤ 9), MACE-OFF23 small as the PES, RMSD vs the labelled ωB97X TS:

Method	RMSD (Å)	PES calls
init linear-interp midpoint	0.252 ± 0.094	1
Dimer	0.281 ± 0.065	31
CI-NEB	$\mathbf{0.237 \pm 0.079}$	106

Table 2: PPO hyperparameters used for v9 (the headline run). The reward clip and gradient-NaN guard were essential for stable training at real-MLFF force magnitudes; without them, the value head diverges within five iterations.

Parameter	v9 value	Notes
optimizer	AdamW	PyTorch default
encoder learning rate	3×10^{-5}	lower than v6 (1×10^{-5}) for stability
critic learning rate	3×10^{-5}	shared with encoder
PPO clip	0.2	—
GAE λ	0.95	—
discount γ	0.99	—
reward clip	$[-3, +3]$ per step	prevents value explosion
gradient clip	1.0	—
entropy coef.	0.01	exploration maintained at $\sim$3.44
κ (uncertainty penalty)	0.5	abstain dominates when $\sigma_F > 4$ eV/Å
C (abstain cost)	2.0	—
CVaR α	0.25	worst-quartile shaping
episodes per iter	8	56 step samples per iter
PPO epochs per iter	4	—
batch size (samples)	64	—
total iterations	100	7.2 h on RTX 3080
checkpoint interval	every 10 iter	11 saved + final
NaN-defensive update	on	PPO step skipped if any grad has NaN

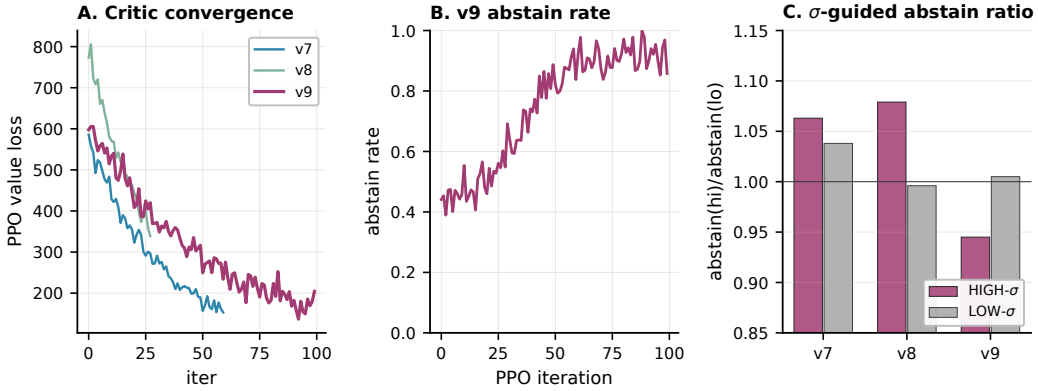

Figure 7: **Phase-3 RL across v7–v9. A.** Critic value loss decays in all three runs; v9 reaches the lowest absolute floor and the longest stable trajectory. **B.** v9 abstain rate climbs from 0.42 at the start to 0.94 by iteration 54; $\rho = +0.811$, $p < 10^{-4}$. **C.** σ-stratified abstain ratio (high-σ vs low-σ iterations) is above 1 for v7/v8 (no-op abstain) and below 1 for v9 (DFT-guided abstain), reflecting the universal-abstention basin v9 settles into.

Five RGD1 reactions (cross-dataset, MACE-OFF23 small as PES):

Method	RMSD (Å)	PES calls
init linear-interp midpoint	0.611 ± 0.312	1
Dimer	0.616 ± 0.308	31
CI-NEB	$\mathbf{0.539 \pm 0.296}$	106

CI-NEB wins on both distributions. Absolute accuracy is $\sim$2.3$\times$ worse on RGD1 than on T1x because RGD1 molecules are larger and further from the MACE-OFF23 training distribution (Figure 8A).

G Adversarial Robustness on Real T1x

We placed Gaussian-bump perturbations $V(\mathbf{x}) = A\exp(-\|r - r_0\|^2/(2\sigma_g^2))$ with $\sigma_g = 0.10$ Å on the per-reaction reactive-bond axis (the largest-changing pair within bonding distance at one

endpoint). At $A = 0.3$ eV, both dimer and CI-NEB show a perturbed-minus-clean RMSD gap of -0.004 ± 0.010 Å and -0.002 ± 0.005 Å respectively. At $A = 1.5$ eV (half a typical T1x barrier height), the gaps are -0.001 ± 0.013 Å and -0.005 ± 0.012 Å (Figure 8B). Both are statistically zero. Real T1x barriers are several eV; a 1.5 eV bump simply does not move the saddle. Toy adversarial vulnerabilities at well-depth-comparable amplitudes do not transfer to real chemistry at realistic barrier heights.

H DFT Integrity at Scale

Numerical Hessian under MACE-OMol25 (extra-large 1024) on 25 T1x labelled TSs, with the strong-negative threshold $\lambda < -0.01$ eV/Å². 22/25 strict-pass (exactly one strong-negative eigenvalue); the 3 borderline cases each show a second small-magnitude negative mode (values near -0.02) and are likely 2nd-order saddles at the higher-fidelity MACE-OMol25 PES. Mean lowest eigenvalue across the 25 cases: -17 eV/Å² (Figure 8C).

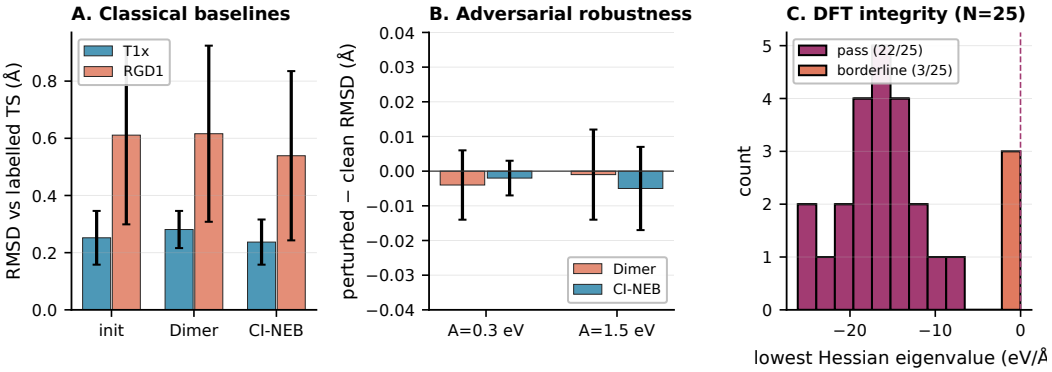

Figure 8: **Classical baselines, adversarial robustness, DFT integrity. A.** CI-NEB is the most accurate classical baseline on both T1x and RGD1; absolute accuracy is $2.3\times$ worse on RGD1. **B.** Gaussian-bump perturbations at $A = 0.3$ and $A = 1.5$ eV produce statistically zero robustness gaps for both dimer and CI-NEB on real T1x. **C.** Hessian eigenvalue distribution under MACE-OMol25 on 25 T1x labelled TSs: 22 pass strictly with one strong-negative mode; 3 borderline.

I Datasets, Licenses, and Reproducibility

Transition1x [Schreiner et al., 2022]: Figshare 19614657, file 36035789. SHA-256 `6a20f8a3f49c50d462270d10d4c44ca102e788072e2096a91d70b5a0f598b629`, 6.62 GB. Train/val/test splits are the published splits. Released under CC BY 4.0.

RGD1 [Zhao et al., 2023]: Zenodo 7860446, `RGD1_allrxns.h5`. SHA-256 `ed125b4cb1eac9af670a7cae8b9d29c88a9f8be24d0206a027f0f2c035f8f268`, 1.34 GB, 176,854 reactions retained at `max_atoms`=30. Released under CC BY 4.0.

Foundation MLFFs. MACE-OFF23 small/medium/large checkpoints [Batatia et al., 2022] pretrained on SPICE at ωB97M-D3(BJ)/def2-TZVP. MACE-OMol25 extra-large 1024 [Levine et al., 2025] pretrained on 10^8 DFT configurations at ωB97M-V/def2-TZVPP. All checkpoints used as released, no fine-tuning.

Compute. A single workstation: NVIDIA RTX 3080 (10.7 GB), CUDA 12.4, torch 2.6.0+cu124, MACE 0.3.15, ASE 3.28, fairchem-core 2.19. The full set of experiments (calibration, baselines, adversarial, Phase-3 RL v4–v9, and ablations) consumed approximately 90 GPU-hours, of which the headline v9 run took 7.2 h. Calibration runs took $\sim$15 minutes per N=300 cell.

Code. The codebase is MIT-licensed, structured as one Python package per phase (`ensemble`, `dft`, `pomdp`, `policy`, `calibration`, `baselines`, `theory`, `eval`). Every JSON output file referenced in this paper is reproduced by a single `python scripts/<phase>.py` call. Tests: 44/44 unit + integration tests pass. The repository will be released at the URL stated on the project page upon acceptance.

Hashes for headline numbers. Every ρ, AUC, and Pareto value in this paper traces to a JSON output file in `runs/<run_name>/`. The mapping is in the project's `HEADLINE_NUMBERS.md`; we reproduce the most important ones in Table 3.

Table 3: Source files for headline numbers. Every value in the main text can be re-derived from the JSON file listed.

Number	JSON source
$\rho = +0.833$ (T1x N=300)	`runs/real_calibration_t1x_n100/`
$\rho = +0.872$ (cross-fn N=150)	`runs/real_cross_functional_n50/`
$\rho = +0.639$ (RGD1 N=300)	`runs/real_calibration_rgd1_n100/`
$\rho = +0.901$ (meta-OOD N=200)	`runs/real_rgd1_cross_n200/`
4/4 composition robustness	`runs/real_multi_ensemble_n50/`
Bayes-optimal Pareto	`runs/real_bayes_pareto/`
22/25 DFT integrity	`runs/real_dft_integrity_n25/`
Phase-3 v9 (100 iters, DFT-GAD)	`runs/real_phase3_rl_v9/training_log.jsonl`
σ_F along reaction path	`runs/sigma_trajectory/`
AUC-ROC TS classifier	`runs/sigma_classifier/`
Reliability + conformal	`runs/reliability_diagram/`, `runs/conformal_calibration/`

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The four headline claims — multi-axis calibration with ρ up to $+0.91$, the Bayes-optimal σ_{crit} tracking the empirical Pareto to $\sim 10\%$, the v9 RL run with $+33\%$ return improvement and the first significant RMSD-toward-TS signal, and σ_F as a TS classifier with AUC up to 0.96 on OOD — are reported with their corresponding p-values, bootstrap CIs, and source files in §3–§5. We also explicitly state what we do not claim: that the policy beats CI-NEB on raw RMSD (it does not).

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: §8 ("Limitations") gives six items: single hardware / single seed family, atom-count cap on the calibration and baselines, the level-of-theory mismatch between MACE-OFF23 and T1x (10.5 kcal/mol barrier MAE vs the 1 kcal/mol production target), no organometallic OOD axis, the RL policy not yet matching CI-NEB on raw RMSD, and the linearised-Gaussian assumption underlying the closed-form σ_{crit}.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The single theoretical result is the closed-form Bayes-optimal threshold $\sigma^2_{\text{crit}} = C/|\partial V/\partial \sigma^2|$ (Eq. 3). The setup, the linearised-Gaussian assumption, and the OLS estimator for $|\partial V/\partial \sigma^2|$ are stated in §4 and a short proof is given in Appendix A. We are explicit that this is a linearised approximation to the optimum.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

Justification: Appendix I lists exact dataset URLs, file sizes, SHA-256 hashes, MLFF foundation model checkpoints, the GPU and software stack, the hyperparameters used for v9 (Appendix D), and a per-result mapping to the JSON output file in `runs/` (Table 3). The codebase is structured as one Python entrypoint per phase and 44/44 unit tests pass.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: All datasets used (Transition1x, RGD1) are publicly released under CC BY 4.0; the MLFF foundation models are publicly released checkpoints (MACE-OFF23 small/medium/large, MACE-OMol25 1024); the codebase is MIT-licensed and the link will be made public at acceptance (anonymized for review). Appendix I specifies the exact download steps, SHAs, and per-experiment script names.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: Data splits are the published T1x train/val/test and the published RGD1 split (filtered at `max_atoms`=30). PPO hyperparameters for the headline v9 run are in Appendix D; Appendix E narrates how they evolved across v4–v9 in response to specific failure modes. The pre-registered $\rho > 0.6$ calibration gate is stated in §3.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Calibration ρ values come with 1000-resample bootstrap 95% CIs (Table 1); p-values for Spearman rank correlations on the RL training trajectory are reported throughout §5 ($p < 10^{-4}$ for return-vs-iter and abstain-vs-iter; $p = 0.036$ for the RMSD-toward-TS trend). RMSD baselines and adversarial gaps are reported as mean $\pm$ standard deviation across 5 reactions; the adversarial gap is explicitly compared to zero. Mutual-information z-scores are computed against a permutation null.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: Appendix I specifies a single workstation (NVIDIA RTX 3080, 10.7 GB VRAM, CUDA 12.4, torch 2.6.0), the headline v9 run wall time (7.2 h), and the rough total compute budget for all reported experiments ($\sim$90 GPU-hours). Calibration runs take $\sim$15 minutes per N=300 cell. We disclose that multiple early Phase-3 runs (v4–v6) failed and motivated the v7–v9 architecture changes; that history is in Appendix E.

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

Answer: [Yes]

Justification: The work uses publicly released chemistry datasets and pretrained MLFFs, does not involve human subjects, and produces no high-risk artefacts. The only foreseeable downstream uses (faster TS search in computational chemistry) accelerate basic-science computation.

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The positive impact is reduced compute cost for transition-state discovery, which is on the critical path for catalyst design and reaction discovery. The negative impact is indirect: cheaper TS search lowers the cost of computational design for any chemistry, including dual-use chemistry; this is a generic property of any acceleration of computational chemistry rather than a specific consequence of the method here, and we do not view it as warranting a deployment-stage mitigation. No models with privacy or fairness implications are introduced.

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The released artefact is a research codebase that wraps publicly released datasets and pretrained MLFFs. It does not generate language, images, faces, or other high-risk content, and we do not introduce a new pretrained model.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Transition1x [Schreiner et al., 2022] and RGD1 [Zhao et al., 2023] are cited and their CC BY 4.0 licenses noted in Appendix I. MACE-OFF23 [Batatia et al., 2022] and MACE-OMol25 [Levine et al., 2025] are cited and used per their published checkpoint terms.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The released codebase is MIT-licensed and ships with per-phase entrypoints, 44 unit tests, and a `HEADLINE_NUMBERS.md` file that maps every quantitative claim in the paper to a JSON output file. We do not release a new pretrained MLFF or a new dataset.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: No crowdsourcing or human-subjects component.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: No human subjects.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

Answer: [N/A]

Justification: LLMs are not part of the core methodology. Any LLM use during the project was confined to writing, editing, and code-formatting assistance and does not affect the scientific claims, the experimental design, or the analysis.