Geometry, Not Energy Surface, Drives the Neutral MLIP–DFT Gap in Atomistic Interaction Surrogates

Rui Ding [*1] **Zixin Ding** [*2] **Rodrigo Pires Ferreira** [*1] **Junhong Chen** [1] **Yuxin Chen** [2]

Abstract

Pretrained universal machine-learning interatomic potentials (MLIPs) are now positioned as high-fidelity surrogate simulators for electronic-structure physics, widely expected to replace explicit quantum chemistry where they cover the chemistry. We show that on non-covalent binding, this expectation is mostly right, but mostly for the wrong reason. We audit the FAIRChem UMA omol head on 5,565 probe–target dimers from 18 benchmarks across nine fidelity arms, without any task-specific fine-tuning. Three findings emerge. First, the neutral MLIP–DFT gap is geometric rather than a learned-surface error: on the neutral aggregate RAPIDS lands at mean absolute error (MAE) 1.37 kcal/mol, matching or beating every explicit DFT single-point fidelity (1.42–1.50) and beating two of three DFT geometry-reoptimized arms; only ωB97M-V GeoSP (1.30 kcal/mol at $\sim 4\times$ the RAPIDS cost) edges RAPIDS by $\sim 5\%$, and MACE-omol and ORB-omol reach the same DFT-SP parity on neutral chemistry. Second, charged systems behave as a qualitatively different regime: MAE jumps to 13.83–27.21 kcal/mol with Spearman $\rho \approx 0$, a DFT single point does not recover it, only DFT geometry reoptimization does, and the failure persists across all three backbones. Third, the residual errors are structured: signed biases cluster by chemistry family across backbones, and a chemistry-aware replay selector (CHEMUCB-8) reaches 1.01 kcal/mol at 52% of the CREST xTB+DFT baseline cost. The audit is delivered through RAPIDS, a training-free inference-time scaffold exposed via a Model Context Protocol (MCP) interface and illustrated on a per- and polyfluoroalkyl substance (PFAS) sensing case study.

[*]Equal contribution [1]Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL, USA [2]Department of Computer Science, University of Chicago, Chicago, IL, USA. Correspondence to: Yuxin Chen <chenyuxin@uchicago.edu>.

Proceedings of the AI4Physics Workshop at the 43rd International Conference on Machine Learning (AI4Physics@ICML 2026), Seoul, South Korea. 2026. Copyright 2026 by the author(s).

1. Introduction

Pretrained universal machine-learning interatomic potentials (MLIPs) such as UMA and MACE deliver order-of-magnitude speedups over explicit quantum chemistry on broad molecular inputs (Levine et al., 2025; Batatia et al., 2022; Wood et al., 2025). Low pretraining loss on large molecular corpora is, however, a statement about average force and energy regression on training-distribution conformers, and gives little direct evidence on whether a downstream user can rank non-covalent complexes within 1–2 kcal/mol across chemically distinct regimes. That distinction matters in chemosensing, host–guest recognition, lead optimization, and porous-material screening, where a 1–2 kcal/mol misorder reassigns which probe, ligand, or host is carried forward to experiment (Barrow et al., 2015; Krämer et al., 2022; Wang et al., 2015; Furukawa et al., 2013).

Existing accurate workflows answer this question one system at a time but not at screening scale. CREST conformer exploration on the GFN2-xTB surface with composite-DFT refinement remains a strong reference for small molecular dimers, yet its wall-clock is dominated by repeated semiempirical gradients and DFT follow-ups (Pracht et al., 2024; Bannwarth et al., 2019; Bursch et al., 2022). Docking engines and learned pose generators are orders of magnitude faster, but their scores are calibrated for protein–ligand or geometric plausibility settings and are not general-purpose electronic binding quantities for chemically diverse non-covalent dimers (Trott & Olson, 2010; Friesner et al., 2004; Corso et al., 2023; Qiao et al., 2024). Pretrained universal MLIPs sit between these regimes in cost, but their evaluation resources are organized around standalone molecules, conformers, and general-purpose energy/force supervision rather than downstream binding tasks (Eastman et al., 2023; Levine et al., 2025). What is missing is a controlled measurement of where a pretrained universal MLIP is reliable inside a realistic screening stack, where it fails, and which intermediate fidelities recover accuracy when it does. Equivalently, this asks whether a pretrained atomistic potential can

serve as an error-controlled surrogate simulator for molecular interaction physics, rather than only as a fast energy regressor.

Non-covalent interaction (NCI) ranking is the right setting for this measurement. It has a natural multi-fidelity ladder spanning MLIP screening, explicit DFT single points, DFT geometry reoptimization, and CREST sampling, paired with published CCSD(T)-quality references and a meaningful neutral-versus-charged regime split. Relative energies often differ by only 1 to 2 kcal/mol, so small systematic errors flip rankings, while the cost gap between the cheapest and most expensive arms spans roughly an order of magnitude, making escalation decisions practically meaningful.

An alternative to fine-tuning a foundation MLIP for downstream tasks (Deng et al., 2024; Kaur et al., 2024; Radova et al., 2025) is to leave the weights untouched and ask how far a structured inference-time scaffold can carry the pretrained model, paralleling how prompt and tool scaffolds adapt large language models to downstream tasks without fine-tuning. We instantiate this route in RAPIDS (Rapid Atomistic Probe–target Interaction Discovery Scaffold): a three-tier orientation search, limited-memory BFGS (LBFGS) relaxation, and a four-guard validation stack wrapped around the pretrained FAIRChem UMA omol head. We benchmark RAPIDS against eight additional fidelities on 5,565 dimers from 18 interaction datasets, producing a $5{,}565 \times 9$ evaluation log that we use as both a benchmark and an offline replay substrate.

Contributions. Concretely, we make the following contributions.

1. **The neutral MLIP-DFT gap is geometry, not energy surface.** On covered neutral chemistry, RAPIDS lands at aggregate mean absolute error (MAE) 1.37 kcal/mol, matching or beating every explicit DFT single-point fidelity (1.42–1.50) and beating two of three DFT geometry-reoptimized arms; only ωB97M-V GeoSP at 1.30 kcal/mol (at $\sim 4\times$ the RAPIDS cost) edges RAPIDS by $\sim 5\%$. A cross-backbone diagnostic on the same RAPIDS-committed geometries with MACE-omol and ORB-omol confirms that all three MLIPs reach DFT-SP parity on neutral chemistry once charged motifs are removed, so the residual gap is geometric rather than a learned-surface deficit.
2. **Charged systems form a qualitatively distinct regime.** MAE jumps to 13.83–27.21 kcal/mol with Spearman $\rho \approx 0$, DFT single points on the MLIP geometry do not recover accuracy, and only DFT geometry reoptimization or CREST sampling does. The failure persists across all three MLIP backbones we test, ruling out a UMA-specific artifact and pointing at a property of how short-range message-passing potentials currently handle charged binding.
3. **Residual errors are structured, and the structure is cheap to exploit.** Signed biases cluster by chemistry family across all three backbones, and replaying 40 allocation heuristics on the same log, a chemistry-aware selector (CHEMUCB-8) reaches 1.01 kcal/mol at 52% of the CREST xTB+DFT baseline cost.
4. **A controlled audit and a reusable evaluation log.** We deliver these findings through a 5,565-dimer, nine-fidelity audit (MLIP screening, three explicit DFT single-points on the MLIP geometry, three DFT geometry-reoptimization rungs, and two CREST-based references) of the FAIRChem UMA omol head with no task-specific fine-tuning, and release the resulting $5{,}565 \times 9$ evaluation log as both a benchmark and an offline replay substrate. The scaffold delivering the audit (RAPIDS) is exposed as a progressively disclosed Skill behind a Model Context Protocol (MCP) interface and illustrated on a per- and polyfluoroalkyl substance (PFAS) sensing case study.

2. Related Work

Interaction screening and conformer search. Ranking non-covalent dimers has been a generate-then-score problem. CREST on the GFN2-xTB surface with composite-DFT refinement sets the accuracy reference but takes hours per dimer (Pracht et al., 2024; Bannwarth et al., 2019; Bursch et al., 2022). Docking engines and diffusion-based pose models are orders of magnitude faster, but their scores target rigid protein-ligand pockets and are not designed for electronic binding across chemically diverse dimers (Trott & Olson, 2010; Friesner et al., 2004; Corso et al., 2023; Qiao et al., 2024). No existing line provides a calibrated low-cost electronic-binding scorer benchmarked on the same task. Appendix A.1 covers application background and Appendix A.2 the broader survey.

Universal MLIPs for molecular interaction scoring. Equivariant neural interatomic potentials such as NequIP, MACE, Allegro, and UMA have pushed molecular surrogate modeling well beyond narrow force-field fitting, especially once coupled to large datasets such as SPICE, OMol25, and related cross-domain corpora (Batzner et al., 2022; Batatia et al., 2022; Musaelian et al., 2023; Eastman et al., 2023; Levine et al., 2025; Wood et al., 2025). What is less developed is a full interaction-screening workflow built around such a model, with explicit pose search, validation, fallback logic, and end-to-end benchmark evaluation on chemically diverse non-covalent dimers. Appendix A.3 gives the longer universal-MLIP background that motivated this framing.

Multi-fidelity allocation and offline replay. Adaptive allocation across cheap and expensive fidelities has a long

literature in multi-armed bandits, contextual bandits, multi-fidelity Bayesian optimization, and offline policy evaluation (Auer et al., 2002; Li et al., 2010; Kandasamy et al., 2016; 2017; Dudík et al., 2011). Those ideas are directly relevant here because our 9-fidelity matrix exposes a decision problem with nontrivial cost-accuracy tradeoffs. The main text uses this perspective operationally rather than as a purely theoretical framing; Appendix A.4 gives the broader literature bridge.

Scientific agents, chemistry tools, and tool interfaces. Chemistry and materials agents have shown that large language models can orchestrate literature search, property prediction, cheminformatics, and simulation tools in longer-horizon workflows (Bran et al., 2024; Boiko et al., 2023; Chen et al., 2025; Kang & Kim, 2024; Zhang et al., 2024; Ding et al., 2026). The Model Context Protocol and progressive-disclosure Skills pattern make it easier to expose such tools to agents without filling the prompt with raw schemas (Anthropic, 2024b; Model Context Protocol Contributors, 2025; Zhang et al., 2025; Anthropic, 2025a). Here these interfaces matter as application-layer packaging for RAPIDS; Appendices A.5 and A.6 give the longer discussion.

3. Methods

We study a real multi-fidelity screening stack in which each of $5{,}565$ retained probe–target dimers from 18 benchmark datasets is logged across nine fidelity arms; the resulting attempted $5{,}565 \times 9$ log records a relaxed binding score, a wall-clock cost, and a status flag per cell, and method-level missing cells (dominated by SCF-convergence failures and a small unsupported-system count) are handled at replay time by the fallback rule of §3.4. The main comparison asks when the cheapest learned surrogate is already useful and when explicit higher-fidelity physics changes the answer. Every fidelity emits a relaxed vacuum electronic binding score $\Delta E_{\mathrm{bind}} = E_{\mathrm{complex}} - E_{\mathrm{probe}} - E_{\mathrm{target}}$ at each pipeline's relaxed geometries, so more negative values indicate more favorable binding; errors are computed against the published benchmark interaction-energy references $\Delta E^{\star}_{\mathrm{int}}$ evaluated at the dataset-provided geometries. Full quantity definitions, benchmark-initialization rules, unsupported-system bookkeeping, and thresholds are in Appendix B.

3.1. Benchmark Setup and Nine-Fidelity Ladder

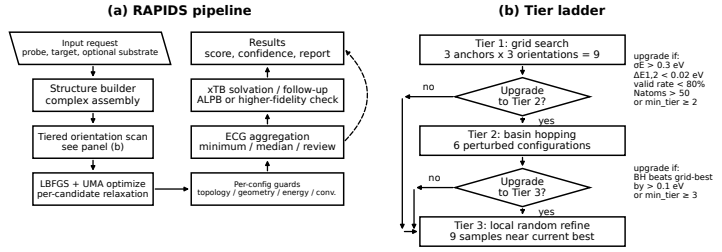
For benchmark evaluation, PubChem retrieval (Kim et al., 2023) is bypassed. Monomer coordinates, fragment charges, and benchmark reference geometries are taken from the dataset records, the monomers are separated from those records, and each pipeline rebuilds its search start from the same monomer pair. For a fixed system and fidelity, isolated probe and target monomers are evaluated or optimized once and then reused across candidate poses of that system. The benchmark families span canonical small-molecule dimers (Řezáč et al., 2015; Rezac et al., 2011; 2012), the Non-Covalent Interactions Atlas H-bond, ionic-H-bond, σ-hole, and London-dispersion benchmarks (Řezáč, 2020a;b; Kříž & Řezáč, 2022; Řezáč, 2022), and the BioFragment Database protein-fragment subsets (Burns et al., 2017). Their published reference values are treated as ground truth throughout: CCSD(T)/CBS-style benchmark interaction energies for the small-molecule and Non-Covalent Interactions Atlas sets, and CCSD(T)-F12 silver-standard interaction energies for BFDb (Řezáč et al., 2015; Rezac et al., 2011; 2012; Řezáč, 2020a;b; Kříž & Řezáč, 2022; Řezáč, 2022; Burns et al., 2017). The detailed benchmark map with counts, chemistry labels, and source references is given in Appendix Table 1.

The nine fidelity arms are RAPIDS; the three fixed-geometry explicit DFT checks PBE-D3BJ/def2-TZVP SP, ωB97X-D3BJ/def2-TZVP SP, and ωB97M-V/def2-TZVP SP; the three geometry-corrected explicit DFT checks PBE-D3BJ GeoSP, ωB97X-D3BJ GeoSP, and ωB97M-V GeoSP; and the two CREST baselines CREST xTB and CREST xTB+DFT (Perdew et al., 1996; Grimme et al., 2010; 2011; Chai & Head-Gordon, 2008; Mardirossian & Head-Gordon, 2016). The SP arms apply explicit def2-TZVP single points on RAPIDS geometries. The GeoSP arms reoptimize the RAPIDS geometry at the same functional with def2-TZVP and then evaluate a def2-TZVPD single point. This ladder is designed so that the RAPIDS score, fixed-geometry explicit DFT checks, geometry-corrected explicit DFT checks, and the full CREST baseline are all evaluated on the same retained system set and can be replayed by downstream allocation policies without launching expensive calculations.

3.2. RAPIDS Scaffold

RAPIDS (Rapid Atomistic Probe–target Interaction Discovery Scaffold) wraps the pretrained FAIRChem Universal Model for Atoms (UMA) omol head (Wood et al., 2025; Levine et al., 2025) in a search-and-validation workflow; we use the UMA weights as released. The scaffold has four parts: a three-tier orientation search (fixed-anchor Tier 1, random-perturbation refinement in Tiers 2–3), LBFGS relaxation with smart continuation, a validation stack that emits confidence labels and fallback recommendations, and an optional xTB plus analytical linearized Poisson–Boltzmann (ALPB) post-correction. Replay and agent-facing interfaces sit on top of the resulting log. Figure 1 compares the guarded workflow to the CREST composite baseline stage by stage.

At the workflow level, RAPIDS starts from the benchmark monomers, evaluates a structured $9+6+9$ orientation ladder on the UMA surface, relaxes each candidate with smart con-



(c) Pipeline comparison

Stage	RAPIDS	CREST + composite-DFT
Search evaluator	UMA omol surrogate	CREST metadynamics plus GFN2-xTB
Pose search	Deterministic 9+6+9 ladder with Tier-2 basin hopping and Tier-3 local refinement	CREST NCI metadynamics in a 6 kcal/mol window
Local optimization	MLIP LBFGS with smart continuation	xTB tight reoptimization, then explicit DFT follow-up
Validation	Topology, geometry, energy, and convergence guards plus ECG	SCF and optimization convergence diagnostics
Explicit DFT rung	Optional follow-up SP or GeoSP on RAPIDS geometries via GPU4PySCF	r2SCAN-3c geometry optimization plus ωB97M-V/def2-TZVPD single point via GPU4PySCF
Reported score	Relaxed vacuum electronic binding score	Relaxed vacuum electronic binding score

Figure 1. RAPIDS architecture and compact pipeline comparison. Panel (a) shows the guarded RAPIDS path: it builds a probe–target(-substrate) complex, executes a tiered orientation scan, relaxes each candidate on the UMA omol surface, applies per-configuration guards together with a scan-level Energy Consistency Guard, and returns confidence-annotated outputs with optional xTB solvation or higher-fidelity follow-up. Panel (b) summarizes the tier-upgrade logic. Panel (c) contrasts RAPIDS with the CREST plus composite-DFT baseline at the stage level.

tinuation, and commits a single pose-ranked relaxed binding score together with confidence metadata. Tier 1 enumerates nine anchor-and-rotation placements, Tier 2 performs six basin-hopping perturbations around the current best candidate, and Tier 3 adds nine local random refinements. The workflow applies four guards, topology, geometry, energy, and convergence, followed by a scan-level Energy Consistency Guard (ECG); all benchmark RAPIDS runs were executed via the MCP `scan_orientations` path so that ECG was active throughout. A key screening rule is the absolute energy guard

$$g_{\mathrm{en}}(r) = \mathbf{1}\Big[\, \frac{|E_{\mathrm{bind}}(r)|}{N_{\mathrm{small}}(r)} > 0.65 \text{ eV/atom} \\ \vee\ |E_{\mathrm{bind}}(r)| > 2.2 \text{ eV} \,\Big],$$

where the 2.2 eV ceiling ($\approx$ 50 kcal/mol) is used as an empirical catastrophic-failure sentinel on the same scale as the 50 kcal/mol evaluation clip. These checks decide when RAPIDS can be trusted as a cheap screening oracle and when a system should be escalated to a higher-fidelity arm; they are not intended to rank close neutral candidates. Exact thresholds, guard logic, ECG formulas, pseudocode, and justification of cutoff are given in Appendix B.1.

3.3. High-Fidelity Baseline and Fidelity Ladder

A strong CREST xTB+DFT benchmark baseline is a cost-controlled CREST plus composite-DFT workflow inspired by Bursch et al. (Bursch et al., 2022). Starting from the same benchmark monomers, it performs CREST non-covalent interaction metadynamics, tight GFN2-xTB reoptimization of the ten lowest xTB candidates, and then forwards the lowest xTB conformer to explicit DFT, using r^{2}SCAN-3c geometry optimization followed by an ωB97M-V/def2-TZVPD single-point evaluation. The isolated monomers are optimized separately with the same method and reused when forming the relaxed binding score. This baseline is intentionally cost-controlled rather than a fully exhaustive DFT reranking protocol. All explicit DFT work uses GPU4PySCF (Sun et al., 2020; Li et al., 2025). The six intermediate DFT arms (3 SP + 3 GeoSP) on top of RAPIDS-committed geometries support the surface/geometry/sampling decomposition of §4.3; full functional, basis, convergence, charge/spin, and monomer-reuse details are in Appendix B.2.

3.4. Offline Replay and Agent-Facing Packaging

Because the same retained systems are evaluated across all nine fidelities, the resulting $5{,}565 \times 9$ matrix can be reused as an offline replay log for cost–accuracy studies. Each row is one probe–target system, each column is one fidelity arm, and each populated cell stores a relaxed binding score, a wall-clock cost, and a status flag. Replaying allocation heuristics and retrospective replay policies over this common log lets us compare fixed fidelities, rule-based escalation strategies, chemistry-aware selectors, and meta-learning selectors without launching new DFT jobs (Auer et al., 2002; Thompson, 1933; Li et al., 2010). The main paper reports 40 replay policies. Rule-based policies and frozen meta-models are the closest to direct deployment; selectors that use retrospective all-arm feedback or online updates should be read as offline replay studies rather than as strict online bandit simulations. A policy π on system s_t returns an ordered sequence $\mathcal{M}_t = (a_{t,1}, \ldots, a_{t,m_t})$, the committed prediction is $a_t^{\dagger} = \mathrm{last_nonNaN}(\mathcal{M}_t)$, and the per-system reward is

$$r_t^{\pi} = -\min\Big(\big|\Delta\hat{E}_{\mathrm{bind}}(s_t, a_t^{\dagger}) - \Delta E_{\mathrm{int}}^{\star}(s_t)\big|,\ 50\,\text{kcal/mol}\Big).$$

Cost accounting, feedback regimes, and the full policy taxonomy are given in Appendix B.5. The replay families include fixed ALWAYS-X baselines, disagreement and progressive-ladder rules, cheap-ensemble selectors, UCB/THOMPSON-style methods, chemistry-aware selectors such as CHEMUCB-8, and stacked or frozen meta-learning models (Mişir & Sebag, 2017; Friedman, 2001). The full policy taxonomy, update rules, and deployability breakdown are given in Appendix B.5.

Agent-facing packaging. The scaffold is also exposed as a progressively disclosed Skill backed by a Model Context Protocol (MCP) server (Anthropic, 2024b; Model Context Protocol Contributors, 2025; Zhang et al., 2025; Anthropic, 2025a), with a Claude Sonnet 4.5 agent (Anthropic, 2025b) permitted to invoke RAPIDS, follow-up DFT/CREST arms, run inspection, and replay-policy queries directly through the tool interface. This packaging does not change the underlying model or thresholds; full MCP and Skill details are in Appendix B.6.

4. Results

4.1. Benchmark setup and the three questions

The benchmark results below use the common $5{,}565$-dimer log defined in §3.2–§3.3. The set comprises A24, S66, and X40 small-molecule dimers (Řezáč et al., 2015; Rezac et al., 2011; 2012), the Non-Covalent Interactions Atlas H-bond, ionic-H-bond, σ-hole, and London-dispersion benchmarks HB375/IHB100, HB300SPX, SH250, and D1200 (Řezáč, 2020a;b; Kříž & Řezáč, 2022; Řezáč, 2022), and the BioFragment Database (BFDb) protein-fragment subsets (Burns et al., 2017). Their published reference levels are CCSD(T)/CBS-style for the small-molecule and Non-Covalent Interactions Atlas sets and CCSD(T)-F12 silver-standard for BFDb. For each retained system we evaluate the same nine fidelity methods: RAPIDS, three DFT single-points on the RAPIDS geometry, three DFT geometry-reoptimization plus single-points from the RAPIDS geometry, CREST xTB, and CREST xTB+DFT. The resulting $5{,}565 \times 9$ log is used throughout the completed-system summaries below; remaining method-level missing cells are dominated by self-consistent-field (SCF)-convergence failures and are handled at replay time by the fallback rule of §3.4. Primary metrics are the capped MAE with per-system errors clipped at 50 kcal/mol and the Spearman rank correlation ρ between each method's relaxed binding score and the published benchmark interaction-energy reference, following the statistics protocol in §3.4. On the neutral aggregate, RAPIDS costs 14.8 h against 125.3 h for a strong CREST xTB+DFT benchmark baseline, Always-CREST xTB+DFT, so the cheapest end of the ladder already lands at 12% of that baseline's cumulative cost. The results answer three questions: **(Q1)** on neutral non-covalent interaction (NCI) dimers and at a fixed geometry, does the UMA omol head usefully replace the explicit ωB97M-V/def2-TZVP single-point rung used in the matrix, and what is the dominant source of the residual RAPIDS↔CREST+DFT benchmark-relative gap; **(Q2)** on neutral chemistry, how does the RAPIDS pipeline compare with each DFT and CREST fidelity in cost–accuracy terms; and **(Q3)** can calibration, allocation heuristics, and replay policies on the offline log improve the neutral cost–accuracy frontier relative to fixed single-fidelity baselines. Charged-system failure and recovery are collected separately in §4.5; the downstream agent-mediated PFAS harness is summarized in §4.7.

4.2. Neutral per-method landscape

We first identify the neutral benchmarks on which RAPIDS is competitive and those on which explicit DFT or CREST-based sampling materially changes the ranking or error profile. Across the 16 neutral benchmarks, no single fidelity uniformly wins; the full per-benchmark MAE/ρ heatmaps are deferred to Appendix Fig. S1. On H-bond sets, GeoSP variants typically take the best MAE, while RAPIDS lands in the 1.4–1.8 kcal/mol range and remains well below CREST xTB on A24/HB375. On halogen, σ-hole, and D1200 dispersion subsets, the winner is almost always one of the three GeoSP variants: CREST xTB is never the per-benchmark winner on a neutral set, and CREST xTB+DFT tops only S66 among the 16 neutral benchmarks. RAPIDS comes within $\sim 20\%$ of the per-benchmark best on D1200_Halogens, X40, and SH250, and is essentially tied with the best CREST variant on D1200_HBCNO. The strongest neutral result for RAPIDS is the BFDb SSI family, where it is the best-MAE method on all four neutral SSI subsets. The BFDb SSI outcome is benchmark-relative. Those reference structures are crystal-constrained fragments rather than in-vacuo minima, so a DFT geometry optimization can move the complex away from the benchmark reference while still improving the underlying electronic description. This also explains the recurring split between MAE and ρ: on several neutral benchmarks, especially the SSI subsets, RAPIDS is closer to the crystal-constrained reference by MAE while DFT-GeoSP variants preserve the ordering of binding strengths more faithfully by ρ. Best-MAE and best-ρ therefore answer different screening questions.

4.3. Neutral decomposition and cost–accuracy frontier

Q1 asks whether replacing the cheap learned arm by explicit physics changes the answer once geometry and sampling are controlled; the intermediate fidelities are designed for this by holding two of three axes (evaluator, geometry, sampler) fixed at a time. Here ωB97M-V/def2-TZVP SP is an explicit single-point on the RAPIDS-committed geometry, while ωB97M-V GeoSP adds a same-functional geometry reoptimization plus a def2-TZVPD single-point. With $\hat{y}_{\mathrm{R}}, \hat{y}_{\mathrm{SP}}, \hat{y}_{\mathrm{Geo}}, \hat{y}_{\mathrm{C}}$ denoting RAPIDS, ωB97M-V SP, ωB97M-V GeoSP, and CREST xTB+DFT scores respectively, the RAPIDS-versus-CREST+DFT gap admits the telescoping decomposition

$$\hat{y}_{\mathrm{R}} - \hat{y}_{\mathrm{C}} = \delta_{\mathrm{surf}} + \delta_{\mathrm{geom}} + \delta_{\mathrm{samp}},$$

with $\delta_{\mathrm{surf}} = \hat{y}_{\mathrm{R}} - \hat{y}_{\mathrm{SP}}$, $\delta_{\mathrm{geom}} = \hat{y}_{\mathrm{SP}} - \hat{y}_{\mathrm{Geo}}$, and $\delta_{\mathrm{samp}} = \hat{y}_{\mathrm{Geo}} - \hat{y}_{\mathrm{C}}$ (per-system definitions and method-string mappings in Appendix B.4).

Two caveats must be named explicitly. First, δ_{geom} folds in the basis-set change from def2-TZVP at the SP stage to def2-TZVPD at the GeoSP stage (§3.3); the effect of this secondary axis is not separately isolated by this decomposition and we report it as a combined geometry-plus-basis contribution. Second, δ_{samp} folds in the optimization-functional change from ωB97M-V at GeoSP to r^{2}SCAN-3c at CREST Stage 3, so the "sampling" label refers to the full stack-difference and does not isolate sampling. We summarize via adjacent MAE-against-reference shifts; cancellation against the reference means these do *not* directly bound $\mathrm{MAE}(|\delta_{\mathrm{surf}}|)$, $\mathrm{MAE}(|\delta_{\mathrm{geom}}|)$, or $\mathrm{MAE}(|\delta_{\mathrm{samp}}|)$. On the neutral aggregate, Always-RAPIDS lands at 1.37 kcal/mol MAE and Always-ωB97M-V/def2-TZVP SP at 1.44 kcal/mol; swapping the UMA surface for an explicit ωB97M-V single-point at the same (RAPIDS) geometry yields no aggregate benchmark-relative improvement and in fact a marginal $+0.07$ kcal/mol regression. Moving to Always-ωB97M-V GeoSP at the same functional drops aggregate MAE to 1.30 kcal/mol, and Always-CREST xTB+DFT sits at 1.80 kcal/mol. The three adjacent ladder shifts on the neutral aggregate are: RAPIDS$\rightarrow$ ωB97M-V/def2-TZVP SP $+0.07$ kcal/mol, ωB97M-V SP$\rightarrow$ ωB97M-V GeoSP -0.14 kcal/mol, and ωB97M-V GeoSP$\rightarrow$CREST xTB+DFT $+0.50$ kcal/mol. The first shift is consistent with a negligible fixed-geometry surrogate gap on covered neutral chemistry; if anything, RAPIDS lands marginally below explicit ωB97M-V SP at the same geometry. The second corresponds to geometry reoptimization at fixed ωB97M-V functional, and remains the larger contributor in magnitude. The third mixes CREST sampling, r^{2}SCAN-3c optimization, and the final SP stack, so we do not read it as a pure sampling contribution. Taken together, the decomposition constrains the Q1 answer on neutral chemistry: within the UMA omol coverage and at a fixed geometry, the benchmark-relative ladder is consistent with a negligible fixed-geometry surrogate contribution and a larger geometry contribution. The charged ladders invert this ordering and are treated separately in §4.5.

Three qualitative zones appear in the static ladder of Fig. 2 (aggregate cost range 14.8–125.3 h across the nine methods). RAPIDS occupies the low-cost corner at 14.8 h (12% of Always-CREST xTB+DFT, $\sim 10\%$ below the fastest DFT-SP baseline), with aggregate MAE (1.37 kcal/mol) at or below the entire DFT-SP cluster (1.42–1.50 kcal/mol) on the neutral aggregate, and within $\sim 5\%$ of the strongest GeoSP arm. DFT-SP methods cluster at intermediate cost and moderate MAE, with the range-separated hybrids (ωB97X-D3BJ and ωB97M-V) slightly ahead of the pure functional. GeoSP sits at 2–4$\times$ DFT-SP cost; only ωB97M-V GeoSP, at 64.8 h and aggregate MAE 1.30 kcal/mol, strictly beats Always-RAPIDS on the neutral aggregate (PBE-D3BJ GeoSP 1.40 and ωB97X-D3BJ GeoSP 1.44 both land above it). CREST xTB+DFT sits at the highest cost and is Pareto-dominated on the neutral aggregate. The neutral static-fidelity picture is therefore consistent with the decomposition: RAPIDS is already a viable low-cost inner-loop predictor for neutral NCI chemistry, and the only fidelity that materially improves on it (at $\sim 4\times$ cost) is ωB97M-V GeoSP — the same rung the downstream PFAS agent commits at in §4.7.

4.4. Why offline replay matters, and what it gains on neutral chemistry

The neutral sections above establish the two ingredients that make replay useful: the low-cost arm is often competitive but not uniformly best, and the static ladder spans a substantial cost from RAPIDS at 14.8 h to Always-CREST xTB+DFT at 125.3 h. The offline-replay protocol (§3.4) then asks what this fully observed log makes possible beyond static method comparison, comparing 40 allocation heuristics or replay policies on the same system stream without launching new expensive calculations. Some policy families use retrospective all-arm updates or post hoc benchmark-reference feedback, so we interpret these curves as offline replay studies; the rule-based and frozen-model methods are the closest to direct deployment.

Figure 3 reports replay trajectories over the same neutral systems. Under the offline-feedback assumptions of this replay study, the strongest frontier improvement comes from CHEMUCB-8. ORACLE attains MAE 0.39 kcal/mol at 52.8 h, and the best replayed selectors approach without reaching it: CHEMUCB-8 leads at 65.1 h / 1.01 kcal/mol, with chemistry-feature and meta-learning variants clustering in the 30–102 h / 1.02–1.11 kcal/mol band (full inventory in Appendix Table 3). Equivalently, CHEMUCB-8 uses $\approx 52\%$ of the Always-CREST xTB+DFT cumulative cost while reducing MAE from 1.80 to 1.01 kcal/mol — a frontier improvement that also exceeds the strongest static GeoSP arm (ωB97M-V GeoSP at 1.30 kcal/mol, 64.8 h). A simple frozen per-benchmark bias-correction baseline (BIASCORR-RAPIDS, 10% labeled calibration slice) reduces neutral MAE from RAPIDS' 1.37 to 1.30 kcal/mol at the same 14.8 h inference cost — a modest but free improvement that we report for completeness; the more substantive frontier-level gains come from the allocation selectors above. The neutral cost–ρ panel broadly tracks the neutral cost–MAE panel: the same chemistry-feature and meta-learning policies that lead MAE also lead ρ (CHEMUCB-8 $\rho = 0.76$, BUCKET PRIOR ADAPTIVE $\rho = 0.76$, UCB $\rho = 0.70$), and ORACLE reaches $\rho = 0.94$. For readability, Fig. 3 shows a representative subset of the 40 replay policies; the full inventory is listed in Appendix Table 3. Additional charged replay curves are deferred to Appendix Fig. S2. System-level error-distribution diagnostics, aggregate Pareto summaries, and family-level policy rankings are deferred to Appendix Figs. S3, S4, S5, and S6.

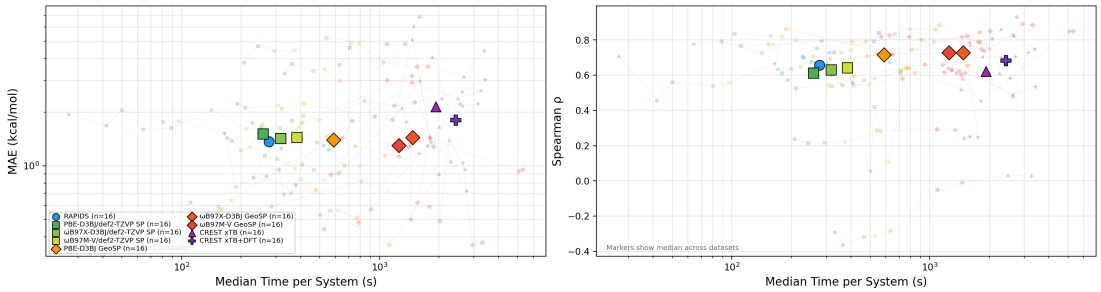


Figure 2. Neutral static-fidelity overview across the 16 neutral benchmarks. Left panel: capped MAE versus median wall-clock time per system, in seconds on a log scale. Right panel: Spearman ρ versus median wall-clock time per system. Faint markers and light gray connectors show benchmark-level values for the nine static fidelities; the larger colored markers show the median across the 16 neutral benchmarks for each fidelity. RAPIDS sits at the low-cost end of the static ladder, while GeoSP and CREST xTB+DFT occupy progressively more expensive rungs.

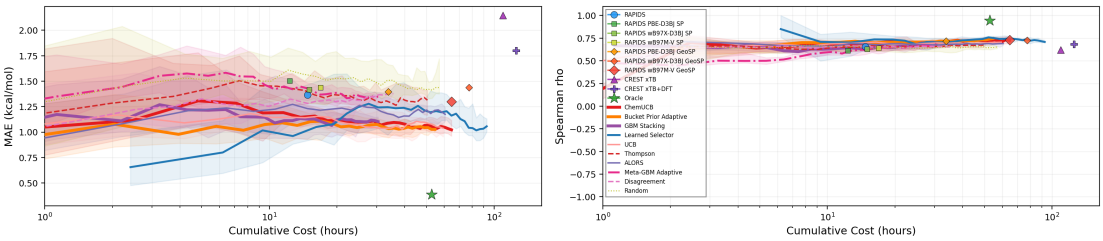


Figure 3. Neutral-only offline-replay trajectories for the 16 neutral benchmarks. The left subplot shows capped MAE versus cumulative cost, in hours on a log scale; the right subplot shows Spearman ρ versus cumulative cost. Dashed lines are fixed single-fidelity baselines; solid lines are representative allocation heuristics and replay policies; shaded bands show the non-parametric $[p_{10}, p_{90}]$ range across ten stream permutations. The best-performing replayed selector, CHEMUCB-8, reaches 65.1 h / 1.01 kcal/mol, which is 52% of the Always-CREST xTB+DFT cumulative cost (125.3 h) while substantially improving MAE (from 1.80 to 1.01 kcal/mol).

4.5. Charged systems: a sharp trust boundary

On charged motifs the same benchmark turns from a speedup story into a reliability story, exposing the first-order trust boundary of the foundation MLIP. Both charged benchmarks (IHB100, 100 ionic hydrogen-bond complexes; BFDb_SSI_charged, 776 salt-bridge and same-charge sidechain pairs) expose a clear limitation of RAPIDS: capped MAE of 13.83 on IHB100 and 27.21 on BFDb_SSI_charged, with Spearman ρ near zero, and adding ωB97M-V/def2-TZVP SP on the RAPIDS geometry does not close the gap. DFT geometry reoptimization is the main recovery mechanism: ωB97M-V GeoSP is the best-MAE method on IHB100 (capped MAE 4.90); on BFDb_SSI_charged, GeoSP helps but is overtaken on MAE by CREST xTB, whose CREST-sampled geometries differ meaningfully from the RAPIDS pose on salt-bridge complexes, while CREST xTB+DFT is slightly worse on MAE but stronger on ρ (per-benchmark MAEs and ρ in Appendix Table 5). This failure pattern is consistent with both missing long-range electrostatics in finite-cutoff MLIPs and thinner pretraining coverage of charged species; we do not disentangle those axes, but the gap between fixed-geometry and geometry-corrected DFT suggests the RAPIDS committed geometry itself is distorted in the charged regime, and the cross-backbone diagnostic in §4.6 shows the failure is only partly backbone-specific. Charged motifs therefore require escalation to a DFT geometry-correction rung or to CREST-based sampling; the validation guards catch catastrophic topology/geometry/energy failures, but charged systems should be escalated by charge/regime membership even when no guard fires (trigger-rate and flagged/unflagged separations in Appendix Tables 7–8; charged diagnostics and case studies in Appendix Figs. S2, S5, S7, and S8, with neutral comparison cases in Appendix Figs. S9–S10).

4.6. Cross-backbone diagnostics and end-to-end replacement

To test whether the failure modes identified above are specific to UMA, we perform a cross-backbone analysis with MACE-omol (Batatia et al., 2022) and ORB-omol (Neumann et al., 2024) at three levels: fixed-geometry single-point scoring on RAPIDS-committed geometries (L1), MLIP-only relaxation of those geometries (L2), and end-to-end replacement (L3) in which the full RAPIDS scaffold runs with the alternative backbone as evaluator. L1/L2 isolate surface bias and relaxation drift under controlled geometry; L3 produces a complete alternative-backbone screening workflow.

Three findings emerge (Appendix Fig. S11). *(i) Chemistry-family-structured signed bias.* Errors are not random noise across UMA, MACE-omol, and ORB-omol; the same hydrogen-bond, σ-hole, halogen, and dispersion families carry consistent signed biases across all three backbones, reinforcing the calibration view of §4.4: the residual surface gap is at least partially *learnable per chemistry family*. *(ii) Paired surface error.* On shared systems, paired MAE against ωB97M-V/def2-TZVP SP is 2.40 kcal/mol for ORB-omol and 2.48 for MACE-omol; the corresponding UMA paired gap is 4.07 kcal/mol, dominated by charged motifs, and once charged motifs are removed all three MLIP backbones lie close to DFT-SP on neutral chemistry. *(iii) Charged behavior splits across backbones at relaxation.* Alternative MLIP scorers reduce but do not eliminate the gap to ωB97M-V GeoSP recovery; MLIP-only relaxation splits across backbones, worsening charged MAE for MACE-omol (capped $\Delta\mathrm{MAE} = +2.30$ kcal/mol) but materially improving it for ORB-omol (-5.15) where it approaches DFT-GeoSP recovery. Explicit DFT geometry correction remains the most reliable across-the-board charged recovery path. Notably, end-to-end L3 replacement reaches near-DFT-GeoSP accuracy on covered chemistry but inherits complementary backbone-specific coverage holes (Appendix Fig. S12); full L3 ranking, BiasCorr cross-backbone universality, and complementary-coverage breakdown are in Appendix C.7. Cross-MLIP disagreement also yields a cheap escalation signal: the standard deviation across predictions correlates with downstream error (Spearman $\rho = 0.655$ overall, 0.841 on charged systems; top-10%-disagreement flag precision 0.734). The alternative backbones thus serve as a small ensemble for calibration and uncertainty-proxy guarding rather than as a Pareto-dominant standalone workflow.

4.7. Agent-mediated PFAS harness

To illustrate downstream agent use of the MCP-exposed scaffold, we ran an agent-mediated PFAS selectivity case study comparing perfluorooctane sulfonic acid (PFOS) and perfluorooctanoic acid (PFOA) against dodecyl sulfate (DDS) and trichloroacetic acid (TCAA), across 18 rollouts under three agent configurations: 6 *single-agent* rollouts, 6 *meta-agent / sub-agent* rollouts, and 6 *meta-agent / sub-executor-agent / sub-critic-agent* rollouts, in which the sub-executor produces candidate probes and the sub-critic scores intermediate selectivity, proposes local analog expansion around emerging winners, and flags underexplored motif families. Fig. 4 shows a representative output: the geometry-reoptimized [2.2.2]-Cryptand quad at the ωB97M-V GeoSP rung, where PFOS (-20.7 kcal/mol) outranks PFOA (-9.6), DDS (-12.3), and TCAA (-11.2). Recovery of the global top-25 PFAS-selective probes scales monotonically with agent configuration: 8/25 (single-agent), 17/25 (meta-agent / sub-agent), 25/25 (meta-agent / sub-executor / sub-critic), with the critic-augmented mode dominated by compact cationic bicyclic and quaternary-ammonium scaffolds. Prompt harnesses, rollout-count details, top-25 probe table, motif analysis, and the full 18-rollout galleries are in Appendix B.6 and Appendix Figs. S13–S30.

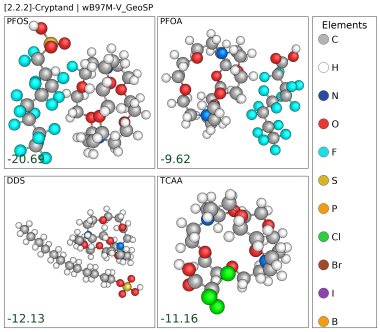


Figure 4. Compact PFAS-harness illustration: geometry-reoptimized [2.2.2]-Cryptand quad at the ωB97M-V GeoSP rung. PFOS is favored over PFOA, DDS, and TCAA.

5. Conclusions

Here we audited a pretrained universal MLIP on non-covalent binding across a 5,565-dimer, nine-fidelity ladder. On covered neutral chemistry, the MLIP is already a strong low-cost screen: on the neutral aggregate RAPIDS lands at or below every DFT single-point fidelity, and the only fidelity that materially improves on it is ωB97M-V GeoSP, by $\sim 5\%$ at roughly $4\times$ the cost. MACE-omol and ORB-omol on the same RAPIDS-committed geometries reach DFT-SP parity once charged motifs are removed, so the neutral MLIP–DFT gap is a geometry gap rather than a learned-surface gap. Charged systems behave differently. MAE jumps by an order of magnitude, Spearman ρ collapses to zero, and DFT single points at the MLIP geometry do not recover accuracy; only DFT geometry reoptimization or CREST sampling does. The same failure repeats across all three backbones, which makes it harder to attribute to UMA's training distribution alone. The errors that remain on neutral chemistry are clustered by chemistry family across backbones, and a chemistry-aware replay selector built on that structure reaches CREST xTB+DFT baseline's cost at roughly half its budget while substantially improving MAE. Our errors are deviations of method-specific relaxed scores against benchmark interaction-energy references rather than frozen-geometry interaction energies; any deployment acting on hits should keep DFT, CREST, or experimental checks in the loop.

References

Agrawal, S. and Goyal, N. Analysis of Thompson Sampling for the Multi-armed Bandit Problem. In *Proceedings of the 25th Annual Conference on Learning Theory (COLT)*, volume 23, pp. 39.1–39.26, 2012. doi: 10.48550/arXiv.1111.1797.

Amezcua, M., Setiadi, J., Ge, Y., and Mobley, D. L. An overview of the sampl8 host-guest binding challenge. *Journal of Computer-Aided Molecular Design*, 36:707–734, 2022. doi: 10.1007/s10822-022-00462-5.

Anthropic. Introducing computer use, a new Claude 3.5 Sonnet, and Claude 3.5 Haiku. https://www.anthropic.com/news/3-5-models-and-computer-use, October 2024a.

Anthropic. Introducing the Model Context Protocol. https://www.anthropic.com/news/model-context-protocol, 2024b. Accessed: 2026-04-18.

Anthropic. Claude can now use tools. https://www.anthropic.com/news/tool-use-ga, May 2024c.

Anthropic. Agent Skills: Overview. https://platform.claude.com/docs/en/agents-and-tools/agent-skills/overview, 2025a.

Anthropic. Introducing Claude Sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5, September 2025b.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2–3):235–256, 2002. doi: 10.1023/A:1013689704352.

Bannwarth, C., Ehlert, S., and Grimme, S. GFN2-xTB – An accurate and broadly parametrized self-consistent tight-binding quantum chemical method with multipole electrostatics and density-dependent dispersion contributions. *J. Chem. Theory Comput.*, 15(3):1652–1671, 2019. doi: 10.1021/acs.jctc.8b01176.

Barrak, A. Traceability and accountability in role-specialized multi-agent LLM pipelines. *arXiv preprint arXiv:2510.07614*, 2025.

Barroso-Luque, L., Shuaibi, M., Fu, X., Wood, B. M., Dzamba, M., Gao, M., Rizvi, A., Zitnick, C. L., and Ulissi, Z. W. Open Materials 2024 (OMat24) Inorganic Materials Dataset and Models. *arXiv preprint*, arXiv:2410.12771, 2024. doi: 10.48550/arXiv.2410.12771.

Barrow, S. J., Kasera, S., Rowland, M. J., del Barrio, J., and Scherman, O. A. Cucurbituril-based molecular recognition. *Chemical Reviews*, 115:12320–12406, 2015. doi: 10.1021/acs.chemrev.5b00341.

Bartók, A. P., Payne, M. C., Kondor, R., and Csányi, G. Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Physical Review Letters*, 104(13):136403, 2010. doi: 10.1103/PhysRevLett.104.136403.

Batatia, I., Kovács, D. P., Simm, G. N. C., Ortner, C., and Csányi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pp. 11423–11436, 2022. doi: 10.48550/arXiv.2206.07697.

Batatia, I., Benner, P., Chiang, Y., Elena, A. M., Kovács, D. P., Riebesell, J., Advincula, X. R., Asta, M., Avaylon, M., Baldwin, W. J., et al. A Foundation Model for Atomistic Materials Chemistry. *arXiv preprint*, arXiv:2401.00096, 2024. doi: 10.48550/arXiv.2401.00096.

Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., Molinari, N., Smidt, T. E., and Kozinsky, B. E(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials. *Nature Communications*, 13(1):2453, 2022. doi: 10.1038/s41467-022-29939-5.

Behler, J. and Parrinello, M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Physical Review Letters*, 98(14):146401, 2007. doi: 10.1103/PhysRevLett.98.146401.

Bhattacharjee, S., Rano, C., Hajra, A., Bandyopadhyay, T., and Biswas, P. K. Regularized machine learning on molecular graph model explains systematic error in DFT enthalpies. *Sci. Rep.*, 11:14978, 2021. doi: 10.1038/s41598-021-93854-w.

Boiko, D. A., MacKnight, R., Kline, B., and Gomes, G. Autonomous Chemical Research with Large Language Models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Bolla, G., Sarma, B., and Nangia, A. K. Crystal engineering of pharmaceutical cocrystals in the discovery and development of improved drugs. *Chemical Reviews*, 122:11514–11603, 2022. doi: 10.1021/acs.chemrev.1c00987.

Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., and Schwaller, P. Augmenting Large Language Models with Chemistry Tools. *Nature Machine Intelligence*, 6(5):525–535, 2024. doi: 10.1038/s42256-024-00832-8.

Burns, L. A., Faver, J. C., Zheng, Z., Marshall, M. S., Smith, D. G. A., Vanommeslaeghe, K., MacKerell, A. D., Merz, K. M., and Sherrill, C. D. The BioFragment Database (BFDb): An open-data platform for computational chemistry analysis of noncovalent interactions. *J. Chem. Phys.*, 147(16):161727, 2017. doi: 10.1063/1.5001028.

Bursch, M., Mewes, J.-M., Hansen, A., and Grimme, S. Best-practice DFT protocols for basic molecular computational chemistry. *Angew. Chem. Int. Ed.*, 61(42): e202205735, 2022. doi: 10.1002/anie.202205735.

Busschaert, N., Caltagirone, C., Van Rossom, W., and Gale, P. A. Applications of supramolecular anion recognition. *Chemical Reviews*, 115:8038–8155, 2015. doi: 10.1021/acs.chemrev.5b00099.

Caldeweyher, E., Ehlert, S., Hansen, A., Neugebauer, H., Spicher, S., Bannwarth, C., and Grimme, S. A generally applicable atomic-charge dependent London dispersion correction. *J. Chem. Phys.*, 150(15):154122, 2019. doi: 10.1063/1.5090222.

Cao, L., Sekutor, M., Zavalij, P. Y., Mlinarić-Majerski, K., Glaser, R., and Isaacs, L. Cucurbit[7]uril·guest pair with an attomolar dissociation constant. *Angew. Chem. Int. Ed.*, 53(4):988–993, 2014. doi: 10.1002/anie.201309635.

Chai, J.-D. and Head-Gordon, M. Long-range corrected hybrid density functionals with damped atom–atom dispersion corrections. *Phys. Chem. Chem. Phys.*, 10(44): 6615–6620, 2008. doi: 10.1039/B810189B.

Chanussot, L., Das, A., Goyal, S., Lavril, T., Shuaibi, M., Rivière, M., Tran, K., Heras-Domingo, J., Ho, C., Hu, W., Palizhati, A., Sriram, A., Wood, B., Yoon, J., Parikh, D., Zitnick, C. L., and Ulissi, Z. The Open Catalyst 2020 (OC20) Dataset and Community Challenges. *ACS Catalysis*, 11(10):6059–6072, 2021. doi: 10.1021/acscatal.0c04525.

Chen, K., Lu, J., Li, J., Yang, X., Du, Y., Wang, K., Shi, Q., Yu, J., Li, L., Qiu, J., Pan, J., Huang, Y., Fang, Q., Heng, P. A., and Chen, G. Chemist-X: Large Language Model-Empowered Agent for Reaction Condition Recommendation in Chemical Synthesis. *arXiv preprint*, arXiv:2311.10776v6, 2025. doi: 10.48550/arXiv.2311.10776.

Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S., and Richardson, D. C. Molprobity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D Biol. Crystallogr.*, 66(1):12–21, 2010. doi: 10.1107/S0907444909042073.

Chung, J. S. and Green, W. H. Machine learning from quantum chemistry to predict experimental solvent effects on reaction rates. *Chem. Sci.*, 15:2059–2075, 2024. doi: 10.1039/D3SC05353A.

Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. DiffDock: Diffusion steps, twists, and turns for molecular docking. In *International Conference on Learning Representations (ICLR)*, 2023. doi: 10.48550/arXiv.2210.01776.

Deng, B., Zhong, P., Jun, K., Riebesell, J., Han, K., Bartel, C. J., and Ceder, G. CHGNet as a Pretrained Universal Neural Network Potential for Charge-Informed Atomistic Modelling. *Nature Machine Intelligence*, 5(9): 1031–1041, 2023. doi: 10.1038/s42256-023-00716-3.

Deng, B., Choi, Y., Zhong, P., Riebesell, J., Anand, S., Li, Z., Jun, K., Persson, K. A., and Ceder, G. Overcoming systematic softening in universal machine learning interatomic potentials by fine-tuning. *arXiv preprint*, arXiv:2405.07105, 2024. doi: 10.48550/arXiv.2405.07105.

Ding, R., Ding, Z., Ferreira, R. P., Chen, Y., and Chen, J. Text-twin-translation (t^{3}): A full-stack machine learning framework for functional material-device systems discovery. In *AI for Accelerated Materials Design - ICLR 2026*, 2026. URL https://openreview.net/forum?id=eqBZpIAHGC.

Donchev, A. G., Taube, A. G., Decolvenaere, E., Hargus, C., McGibbon, R. T., Law, K.-H., Gregersen, B. A., Li, J.-L., Palmo, K., Siva, K., Bergdorf, M., Klepeis, J. L., and Shaw, D. E. Quantum Chemical Benchmark Databases of Gold-Standard Dimer Interaction Energies. *Scientific Data*, 8(1):55, 2021. doi: 10.1038/s41597-021-00833-x.

Dudík, M., Langford, J., and Li, L. Doubly Robust Policy Evaluation and Learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 1097–1104, 2011. doi: 10.48550/arXiv.1103.4601.

Dudík, M., Erhan, D., Langford, J., and Li, L. Doubly Robust Policy Evaluation and Optimization. *Statistical Science*, 29(4):485–511, 2014. doi: 10.1214/14-STS500.

Eastman, P., Behara, P. K., Dotson, D. L., Galvelis, R., Herr, J. E., Horton, J. T., Mao, Y., Chodera, J. D., Pritchard, B. P., Wang, Y., De Fabritiis, G., and Markland, T. E. 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.

Ehlert, S., Stahn, M., Spicher, S., and Grimme, S. Robust and efficient implicit solvation model for fast semiempirical methods. *J. Chem. Theory Comput.*, 17(7):4250–4261, 2021. doi: 10.1021/acs.jctc.1c00471.

Falkner, S., Klein, A., and Hutter, F. BOHB: Robust and Efficient Hyperparameter Optimization at Scale. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, volume 80, pp. 1437–1446, 2018. doi: 10.48550/arXiv.1807.01774.

Folch, J. P., Lee, R. M., Shafei, B., Walz, D., Tsay, C., van der Wilk, M., and Misener, R. Combining Multi-Fidelity Modelling and Asynchronous Batch Bayesian Optimization. *Computers & Chemical Engineering*, 172:108194, 2023. doi: 10.1016/j.compchemeng.2023.108194.

Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., and Shenkin, P. S. Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *J. Med. Chem.*, 47(7):1739–1749, 2004. doi: 10.1021/jm0306430.

Furukawa, H., Cordova, K. E., O'Keeffe, M., and Yaghi, O. M. The chemistry and applications of metal-organic frameworks. *Science*, 341:1230444, 2013. doi: 10.1126/science.1230444.

Gao, B., Huang, Y., Liu, Y., Xie, W., Ma, W.-Y., Zhang, Y.-Q., and Lan, Y. PharmAgents: Building a Virtual Pharma with Large Language Model Agents. *arXiv preprint*, arXiv:2503.22164, 2025. doi: 10.48550/arXiv.2503.22164.

Grimme, S. Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations. *J. Chem. Theory Comput.*, 15(5):2847–2862, 2019. doi: 10.1021/acs.jctc.9b00143.

Grimme, S., Antony, J., Ehrlich, S., and Krieg, H. A consistent and accurate *ab initio* parametrization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu. *J. Chem. Phys.*, 132(15):154104, 2010. doi: 10.1063/1.3382344.

Grimme, S., Ehrlich, S., and Goerigk, L. Effect of the damping function in dispersion corrected density functional theory. *J. Comput. Chem.*, 32(7):1456–1465, 2011. doi: 10.1002/jcc.21759.

Grimme, S., Bohle, F., Hansen, A., Pracht, P., Spicher, S., and Stahn, M. Efficient quantum chemical calculation of structure ensembles and free energies for nonrigid molecules. *J. Phys. Chem. A*, 125(19):4039–4054, 2021a. doi: 10.1021/acs.jpca.1c00971.

Grimme, S., Hansen, A., Ehlert, S., and Mewes, J.-M. r^2SCAN-3c: A "Swiss army knife" composite electronic-structure method. *J. Chem. Phys.*, 154(6):064103, 2021b. doi: 10.1063/5.0040021.

Hennig, P. and Schuler, C. J. Entropy Search for Information-Efficient Global Optimization. *Journal of Machine Learning Research*, 13(57):1809–1837, 2012. doi: 10.48550/arXiv.1112.1217.

Huang, D., Allen, T. T., Notz, W. I., and Miller, R. A. Sequential Kriging Optimization Using Multiple-Fidelity Evaluations. *Structural and Multidisciplinary Optimization*, 32(5):369–382, 2006. doi: 10.1007/s00158-005-0587-0.

Jamieson, K. and Talwalkar, A. Non-stochastic Best Arm Identification and Hyperparameter Optimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 51, pp. 240–248, 2016. doi: 10.48550/arXiv.1502.07943.

Jia, S., Zhang, C., and Fung, V. LLMatDesign: Autonomous Materials Discovery with Large Language Models. *arXiv preprint*, arXiv:2406.13163, 2024. doi: 10.48550/arXiv.2406.13163.

Jing, B., Corso, G., Chang, J., Barzilay, R., and Jaakkola, T. Torsional diffusion for molecular conformer generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 24240–24253, 2022. doi: 10.48550/arXiv.2206.01729.

Kandasamy, K., Dasarathy, G., Oliva, J. B., Schneider, J., and Póczos, B. Gaussian Process Bandit Optimisation with Multi-fidelity Evaluations. In *Advances in Neural Information Processing Systems 29 (NeurIPS)*, pp. 992–1000, 2016. doi: 10.48550/arXiv.1603.06288.

Kandasamy, K., Dasarathy, G., Schneider, J., and Póczos, B. Multi-fidelity Bayesian Optimisation with Continuous Approximations. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70, pp. 1799–1808, 2017. doi: 10.48550/arXiv.1703.06240.

Kang, Y. and Kim, J. ChatMOF: An Artificial Intelligence System for Predicting and Generating Metal-Organic Frameworks Using Large Language Models. *Nature Communications*, 15(1):4705, 2024. doi: 10.1038/s41467-024-48998-4.

Kaur, H., Della Pia, F., Batatia, I., Advincula, X. R., Shi, B. X., Lan, J., Csányi, G., Michaelides, A., and Kapil, V. Data-efficient fine-tuning of foundational models for first-principles quality sublimation enthalpies. *arXiv preprint*, arXiv:2405.20217, 2024. doi: 10.48550/arXiv.2405.20217.

Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J., and Bolton, E. E. Pubchem 2023 update. *Nucleic Acids Res.*, 51(D1):D1373–D1380, 2023. doi: 10.1093/nar/gkac956.

Kovács, D. P., Moore, J. H., Browning, N. J., Batatia, I., Horton, J. T., Pu, Y., Kapil, V., Witt, W. C., Magdău, I.-B., Cole, D. J., and Csányi, G. 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.

Krämer, J., Kang, R., Grimm, L. M., De Cola, L., Picchetti, P., and Biedermann, F. Molecular probes, chemosensors, and nanosensors for optical detection of biorelevant molecules and ions in aqueous media and biofluids. *Chemical Reviews*, 122:3459–3636, 2022. doi: 10.1021/acs.chemrev.1c00746.

Kreno, L. E., Leong, K., Farha, O. K., Allendorf, M., Van Duyne, R. P., and Hupp, J. T. Metal–organic framework materials as chemical sensors. *Chemical Reviews*, 112:1105–1125, 2012. doi: 10.1021/cr200324t.

Kříž, K. and Řezáč, J. Non-covalent interactions atlas benchmark data sets 4: σ-hole interactions. *Phys. Chem. Chem. Phys.*, 24:14794–14804, 2022. doi: 10.1039/D2CP01600A.

Kruse, H. and Grimme, S. A geometrical correction for the inter- and intra-molecular basis set superposition error in hartree–fock and density functional theory calculations for large systems. *J. Chem. Phys.*, 136(15):154101, 2012. doi: 10.1063/1.3700154.

Lai, T. L. and Robbins, H. Asymptotically Efficient Adaptive Allocation Rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. doi: 10.1016/0196-8858(85)90002-8.

Laio, A. and Parrinello, M. Escaping free-energy minima. *Proc. Natl. Acad. Sci. U.S.A.*, 99(20):12562–12566, 2002. doi: 10.1073/pnas.202427399.

LangChain. LangGraph: Stateful, multi-actor applications with LLMs. https://github.com/langchain-ai/langgraph, 2024.

Larsen, A. H., Mortensen, J. J., Blomqvist, J., Castelli, I. E., Christensen, R., Dułak, M., Friis, J., Groves, M. N., Hammer, B., Hargus, C., et al. The atomic simulation environment—a python library for working with atoms. *J. Phys.: Condens. Matter*, 29(27):273002, 2017. doi: 10.1088/1361-648X/aa680e.

Levine, D. S., Shuaibi, M., Spotte-Smith, E. W. C., Taylor, M. G., Hasyim, M. R., Michel, K., Batatia, I., Csányi, G., Dzamba, M., Eastman, P., Frey, N. C., Fu, X., Gharakhanyan, V., Krishnapriyan, A. S., Rackers, J. A., Raja, S., Rizvi, A., Rosen, A. S., Ulissi, Z., Vargas, S., Zitnick, C. L., Blau, S. M., and Wood, B. M. The Open Molecules 2025 (OMol25) Dataset, Evaluations, and Models. *arXiv preprint*, arXiv:2505.08762, 2025. doi: 10.48550/arXiv.2505.08762.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A Contextual-Bandit Approach to Personalized News Article Recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pp. 661–670, 2010. doi: 10.1145/1772690.1772758.

Li, L., Chu, W., Langford, J., and Wang, X. Unbiased Offline Evaluation of Contextual-Bandit-Based News Article Recommendation Algorithms. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, pp. 297–306, 2011. doi: 10.1145/1935826.1935878.

Li, L., Jamieson, K., DeSalvo, G., Rostamizadeh, A., and Talwalkar, A. Hyperband: A Novel Bandit-Based Approach to Hyperparameter Optimization. *Journal of Machine Learning Research*, 18(185):1–52, 2018. doi: 10.48550/arXiv.1603.06560.

Li, R., Sun, Q., Zhang, X., and Chan, G. K.-L. Introducing GPU acceleration into the python-based simulations of chemistry framework. *J. Phys. Chem. A*, 129(5):1459–1468, 2025. doi: 10.1021/acs.jpca.4c05876.

Liao, Y.-L., Wood, B., Das, A., and Smidt, T. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. In *International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2306.12059.

Ma, Y., Gou, Z., Hao, J., Xu, R., Wang, S., Pan, L., Yang, Y., Cao, Y., Sun, A., Awadalla, H., and Chen, W. SciAgent: Tool-Augmented Language Models for Scientific Reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2024. doi: 10.48550/arXiv.2402.11451.

Mardirossian, N. and Head-Gordon, M. ωB97M-V: A combinatorially optimized, range-separated hybrid, meta-GGA density functional with VV10 nonlocal correlation. *J. Chem. Phys.*, 144(21):214110, 2016. doi: 10.1063/1.4952647.

McNaughton, A. D., Ramalaxmi, G. K. S., Kruel, A., Knutson, C. R., Varikoti, R. A., and Kumar, N. CACTUS: Chemistry Agent Connecting Tool Usage to Science. *ACS Omega*, 9(46):46563–46573, 2024. doi: 10.1021/acsomega.4c08408.

McNutt, A. T., Francoeur, P., Aggarwal, R., Masuda, T., Meli, R., Ragoza, M., Sunseri, J., and Koes, D. R. GNINA 1.0: Molecular docking with deep learning. *J. Cheminformatics*, 13(1):43, 2021. doi: 10.1186/s13321-021-00522-2.

Meng, W.-Q., Sedgwick, A. C., Kwon, N., Sun, M., Xiao, K., He, X.-P., Anslyn, E. V., James, T. D., and Yoon, J. Fluorescent probes for the detection of chemical warfare agents. *Chemical Society Reviews*, 52:601–662, 2023. doi: 10.1039/D2CS00650B.

Microsoft. Semantic Kernel: A lightweight SDK for integrating LLMs with conventional programming languages. https://learn.microsoft.com/en-us/semantic-kernel/overview/, 2024.

Microsoft. Beyond the tools: Adding MCP in VS Code. https://code.visualstudio.com/blogs/2025/05/12/agent-mode-meets-mcp, May 2025.

Mişir, M. and Sebag, M. ALORS: An algorithm recommender system. *Artif. Intell.*, 244:291–314, 2017. doi: 10.1016/j.artint.2016.12.001.

Model Context Protocol Contributors. Model Context Protocol: Specification and documentation. https://modelcontextprotocol.io, 2025.

Moura, J. a. CrewAI: Framework for orchestrating role-playing, autonomous AI agents. https://github.com/crewAIInc/crewAI, 2024.

Musaelian, A., Batzner, S., Johansson, A., Sun, L., Owen, C. J., Kornbluth, M., and Kozinsky, B. Learning Local Equivariant Representations for Large-Scale Atomistic Dynamics. *Nature Communications*, 14(1):579, 2023. doi: 10.1038/s41467-023-36329-y.

Neumann, M., Gin, J., Rhodes, B., Bennett, S., Li, Z., Choubisa, H., Hussey, A., and Godwin, J. Orb: A Fast, Scalable Neural Network Potential. *arXiv preprint*, arXiv:2410.22570, 2024. doi: 10.48550/arXiv.2410.22570.

OpenAI. Function calling and other API updates. https://openai.com/index/function-calling-and-other-api-updates/, 2023.

Patil, S. G., Zhang, T., Wang, X., and Gonzalez, J. E. Gorilla: Large language model connected with massive APIs. *arXiv preprint arXiv:2305.15334*, 2023.

Perdew, J. P., Burke, K., and Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.*, 77(18):3865–3868, 1996. doi: 10.1103/PhysRevLett.77.3865.

Poloczek, M., Wang, J., and Frazier, P. I. Multi-Information Source Optimization. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, pp. 4288–4298, 2017. doi: 10.48550/arXiv.1603.00389.

Pracht, P., Bohle, F., Hansen, A., and Grimme, S. CREST – A program for the exploration of low-energy molecular chemical space. *J. Chem. Phys.*, 160(11):114110, 2024. doi: 10.1063/5.0197592.

Precup, D., Sutton, R. S., and Singh, S. P. Eligibility Traces for Off-Policy Policy Evaluation. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pp. 759–766, 2000. doi: 10.5555/645529.658134.

Putta, P., Mills, E., Garg, N., Motwani, S., Finn, C., Garg, D., and Rafailov, R. Agent Q: Advanced Reasoning and Learning for Autonomous AI Agents. *arXiv preprint*, arXiv:2408.07199, 2024. doi: 10.48550/arXiv.2408.07199.

Qiao, Z., Nie, W., Vahdat, A., Miller, T. F., and Anandkumar, A. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nat. Mach. Intell.*, 6:195–208, 2024. doi: 10.1038/s42256-024-00792-z.

Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Hong, L., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., and Sun, M. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *International Conference on Learning Representations (ICLR)*, 2024.

Radova, M., Stenczel, T. K., Elena, A. M., Best, J. P., Brewick, P. T., Lazić, P., and Bartók, A. P. Fine-tuning foundation models of materials interatomic potentials with frozen transfer learning. *npj Computational Materials*, 2025. doi: 10.1038/s41524-025-01727-x.

Rappoport, D. and Furche, F. Property-optimized Gaussian basis sets for molecular response calculations. *J. Chem. Phys.*, 133(13):134105, 2010. doi: 10.1063/1.3484283.

Řezáč, J. Non-covalent interactions atlas benchmark data sets: Hydrogen bonding. *J. Chem. Theory Comput.*, 16(4):2355–2368, 2020a. doi: 10.1021/acs.jctc.9b01265.

Řezáč, J. Non-covalent interactions atlas benchmark data sets 2: Hydrogen bonding in an extended chemical space. *J. Chem. Theory Comput.*, 16(10):6305–6316, 2020b. doi: 10.1021/acs.jctc.0c00715.

Řezáč, J. Non-covalent interactions atlas benchmark data sets 5: London dispersion in an extended chemical space. *Phys. Chem. Chem. Phys.*, 24:14780–14793, 2022. doi: 10.1039/D2CP01602H.

Rezac, J., Riley, K. E., and Hobza, P. S66: A well-balanced database of benchmark interaction energies relevant to biomolecular structures. *J. Chem. Theory Comput.*, 7(8): 2427–2438, 2011. doi: 10.1021/ct2002946.

Rezac, J., Riley, K. E., and Hobza, P. Benchmark calculations of noncovalent interactions of halogenated molecules. *J. Chem. Theory Comput.*, 8(11):4285–4292, 2012. doi: 10.1021/ct300647k.

Řezáč, J., Dubecký, M., Jurečka, P., and Hobza, P. Extensions and applications of the A24 data set of accurate interaction energies. *Phys. Chem. Chem. Phys.*, 17:19268–19277, 2015. doi: 10.1039/C5CP03151F.

Riniker, S. and Landrum, G. A. Better informed distance geometry: Using what we know to improve conformation generation. *J. Chem. Inf. Model.*, 55(12):2562–2574, 2015. doi: 10.1021/acs.jcim.5b00654.

Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Hambro, E., Zettlemoyer, L., Cancedda, N., and Scialom, T. Toolformer: Language Models Can Teach Themselves to Use Tools. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023. doi: 10.48550/arXiv.2302.04761.

Sempionatto, J. R., Lasalde-Ramirez, J. A., Mahato, K., Wang, J., and Gao, W. Wearable chemical sensors for biomarker discovery in the omics era. *Nature Reviews Chemistry*, 6:899–915, 2022. doi: 10.1038/s41570-022-00439-w.

Shetty, D., Khedkar, J. K., Park, K. M., Kim, K., Koiry, D., Biedermann, F., and Nau, W. M. Can we beat the biotin–avidin pair?: cucurbit[7]uril-based ultrahigh affinity host–guest complexes and their applications. *Chem. Soc. Rev.*, 44:8747–8761, 2015. doi: 10.1039/C5CS00631G.

Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., and Yao, S. Reflexion: Language agents with verbal reinforcement learning. *arXiv preprint arXiv:2303.11366*, 2023.

Smith, J. S., Nebgen, B. T., Zubatyuk, R., Lubbers, N., Devereux, C., Barros, K., Tretiak, S., Isayev, O., and Roitberg, A. E. Approaching Coupled Cluster Accuracy with a General-Purpose Neural Network Potential Through Transfer Learning. *Nature Communications*, 10 (1):2903, 2019. doi: 10.1038/s41467-019-10827-4.

Spicher, S. and Grimme, S. Robust atomistic modeling of materials, organometallic, and biochemical systems. *Angew. Chem. Int. Ed.*, 59(36):15665–15673, 2020. doi: 10.1002/anie.202004239.

Sugita, Y. and Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.*, 314(1–2):141–151, 1999. doi: 10.1016/S0009-2614(99)01123-9.

Sun, Q., Zhang, X., Banerjee, S., Bao, P., Barbry, M., Blunt, N. S., Bogdanov, N. A., Booth, G. H., Chen, J., Cui, Z.-H., Eriksen, J. J., Gao, Y., Guo, S., Hermann, J., Hermes, M. R., Koh, K., Koval, P., Lehtola, S., Li, Z., Liu, J., Mardirossian, N., McClain, J. D., Motta, M., Mussard, B., Pham, H. Q., Pulkin, A., Purwanto, W., Robinson, P. J., Ronca, E., Sayfutyarova, E. R., Scheurer, M., Schurkus, H. F., Smith, J. E. T., Sun, C., Sun, S.-N., Upadhyay, S., Wagner, L. K., Wang, X., White, A., Whitfield, J. D., Williamson, M. J., Wouters, S., Yang, J., Yu, J. M., Zhu, T., Berkelbach, T. C., Sharma, S., Sokolov, A. Y., and Chan, G. K.-L. Recent developments in the PySCF program package. *J. Chem. Phys.*, 153(2):024109, 2020. doi: 10.1063/5.0006074.

Swaminathan, A. and Joachims, T. Batch Learning from Logged Bandit Feedback through Counterfactual Risk Minimization. *Journal of Machine Learning Research*, 16 (52):1731–1755, 2015. doi: 10.5555/2789272.2886805.

Swersky, K., Snoek, J., and Adams, R. P. Freeze-Thaw Bayesian Optimization. *arXiv preprint*, arXiv:1406.3896: 1–12, 2014. doi: 10.48550/arXiv.1406.3896.

Tan, A. J., Descoteaux, M., Kotak, M., Nascimento, G. S., Kavanagh, S. R., Zichi, A., Wang, A., Saluja, K., Hu, H., Smidt, T. E., Johansson, A., Witt, W. C., Kozinsky, B., and Musaelian, A. High-performance training and inference for deep equivariant interatomic potentials. *arXiv preprint*, arXiv:2504.16068, 2025. doi: 10.48550/arXiv.2504.16068.

Tang, X., Hu, T., Ye, M., Shao, Y., Yin, X., Ouyang, S., Zhou, W., Lu, P., Zhang, Z., Zhao, Y., Cohan, A., and Gerstein, M. ChemAgent: Self-Updating Library in Large Language Models Improves Chemical Reasoning. In *International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2501.06590.

Thompson, W. R. On the Likelihood That One Unknown Probability Exceeds Another in View of the Evidence of Two Samples. *Biometrika*, 25(3–4):285–294, 1933. doi: 10.1093/biomet/25.3-4.285.

Trott, O. and Olson, A. J. AutoDock Vina: Improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.*, 31(2):455–461, 2010. doi: 10.1002/jcc.21334.

Wang, L., Wu, Y., Deng, Y., Kim, B., Pierce, L., Krilov, G., Lupyan, D., Robinson, S., Dahlgren, M. K., Greenwood,

J., Romero, D. L., Masse, C., Knight, J. L., Steinbrecher, T., Beuming, T., Damm, W., Harder, E., Sherman, W., Brewer, M., Wester, R., Murcko, M., Frye, L., Farid, R., Lin, T., Mobley, D. L., Jorgensen, W. L., Berne, B. J., Friesner, R. A., and Abel, R. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society*, 137:2695–2703, 2015. doi: 10.1021/ja512751q.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E. H., Narang, S., Chowdhery, A., and Zhou, D. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *International Conference on Learning Representations (ICLR)*, 2023. doi: 10.48550/arXiv.2203.11171.

Weigend, F. and Ahlrichs, R. Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: Design and assessment of accuracy. *Phys. Chem. Chem. Phys.*, 7(18):3297–3305, 2005. doi: 10.1039/B508541A.

Wood, B. M., Dzamba, M., Fu, X., Gao, M., Shuaibi, M., Barroso-Luque, L., Abdelmaqsoud, K., Gharakhanyan, V., Kitchin, J. R., Levine, D. S., Michel, K., Sriram, A., Cohen, T., Das, A., Rizvi, A., Sahoo, S. J., Ulissi, Z. W., and Zitnick, C. L. UMA: A Family of Universal Models for Atoms. *arXiv preprint*, arXiv:2506.23971, 2025. doi: 10.48550/arXiv.2506.23971.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., Awadallah, A. H., White, R. W., Burger, D., and Wang, C. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. In *Conference on Language Modeling (COLM)*, 2024.

Xu, M., Yu, L., Song, Y., Shi, C., Ermon, S., and Tang, J. GeoDiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations (ICLR)*, 2022. doi: 10.48550/arXiv.2203.02923.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., and Narasimhan, K. Tree of Thoughts: Deliberate Problem Solving with Large Language Models. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, 2023a. doi: 10.48550/arXiv.2305.10601.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations (ICLR)*, 2023b. doi: 10.48550/arXiv.2210.03629.

Ye, Z., Liu, Y., and Li, Q. Recent progress in smart electronic nose technologies enabled with machine learning methods. *Sensors*, 21:7620, 2021. doi: 10.3390/s21227620.

Yu, B., Baker, F. N., Chen, Z., Herb, G., Gou, B., Adu-Ampratwum, D., Ning, X., and Sun, H. Tooling or Not Tooling? The Impact of Tools on Language Agents for Chemistry Problem Solving. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 7635–7655, 2025. doi: 10.18653/v1/2025.findings-naacl.424.

Zhang, B., Lazuka, K., and Murag, M. Equipping agents for the real world with Agent Skills. https://www.anthropic.com/engineering/equipping-agents-for-the-real-world-with-agent-sk October 2025.

Zhang, H., Song, Y., Hou, Z., Miret, S., and Liu, B. HoneyComb: A Flexible LLM-Based Agent System for Materials Science. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024. doi: 10.48550/arXiv.2409.00135.

A. Additional Related-Work Background

This appendix collects the longer background discussions that support the main-text Introduction and Related Work sections. It retains the fuller application motivation, search-and-score pipeline literature, universal-MLIP background, multi-fidelity decision literature, chemistry-agent survey, and MCP-plus-Skills interface discussion that were compressed in the main paper for readability.

A.1. Why non-covalent molecular binding matters

RAPIDS targets a concrete computational task: for a probe–target pair, predict a pose-ranked relaxed vacuum electronic binding score

$$\Delta E_{\text{bind}} = E_{PT}(R_{PT}^{\text{opt}}) - E_P(R_P^{\text{opt}}) - E_T(R_T^{\text{opt}})$$

on a FAIRChem UMA omol surrogate trained at the ωB97M-V/def2-TZVPD level, with an optional Analytical Linearized Poisson–Boltzmann (ALPB) implicit-water single-point correction. The quantity we study is a relative energetic ordering across candidate non-covalent complexes. It sits apart from binding free energies, docking scores, and kinetic observables. Many downstream applications depend on resolving such orderings when competing poses are separated by only 1–2 kcal/mol, at throughputs well beyond first-principles screening.

A.1.1. Chemical sensors and biosensors as an application anchor

Synthetic chemosensors and biosensors built around macrocyclic hosts (cyclodextrins, calixarenes, cucurbit[n]urils, pillararenes) transduce host–guest recognition into optical or electrochemical signals, and small relative energetic differences among competing host–analyte complexes often track sensor selectivity against structurally similar interferents (Krämer et al., 2022). The cucurbituril family illustrates how tightly tuned this ordering can be: CB[7] and CB[8] reach monovalent association constants up to 10^{17} M^{-1} in water via ion–dipole contacts at the portals, hydrophobic inclusion, and release of high-energy cavity water (Barrow et al., 2015). Anion sensing is even more constrained because hydration enthalpies exceed most receptor–anion interaction energies, forcing designs that rely on directional C–H$\cdots$X$^-$, halogen-bond, and cooperative hydrogen-bond contacts (Busschaert et al., 2015).

Device-level platforms inherit the same relative-ordering problem. Luminescent metal–organic frameworks embed host–guest chemistry in a porous scaffold whose emission shifts with guest inclusion, enabling discrimination among volatile organics, heavy-metal ions, and explosives (Kreno et al., 2012). Broadly responsive chemiresistive elements recover selectivity statistically through electronic-nose arrays whose response patterns are classified by machine learning models trained on large adsorption landscapes (Ye et al., 2021). Wearable and point-of-care devices extend the same chemistry to sweat, tears, interstitial fluid, and breath, requiring receptors that separate biomarkers from structurally similar metabolites at sub-micromolar concentration (Sempionatto et al., 2022), and probes for nitroaromatic explosives, nerve-agent simulants, and chemical-warfare agents depend on turn-on/turn-off events driven by favorable non-covalent contacts that outcompete environmental interferents (Meng et al., 2023). Kinetics, transduction, surface coverage, and matrix effects enter the observed signal in each case; what the computation is asked to supply is a reliable ranking of candidate probe–analyte interaction energies under small gaps and across flexible partners.

A.1.2. The same pattern across adjacent domains

The same pattern recurs outside sensing. Structure-based drug discovery advances when congeneric ligands are reordered by a few kcal/mol against the target pocket and off-targets; alchemical free-energy protocols tightened this ranking but still inherit the underlying non-covalent energetics (Wang et al., 2015), and macrocyclic host–guest blind challenges continue to stress-test the same physics (Amezcua et al., 2022). In porous materials, gas storage, isomer separation, and catalytic selectivity are decided by $\lesssim$2 kcal/mol differences between competing guest binding modes (Furukawa et al., 2013), and pharmaceutical-cocrystal formation is controlled by relative lattice energies across competing polymorphs (Bolla et al., 2022). The question handed to the energy model often reduces to the same ranking subproblem: order candidate complexes under small gaps at screening throughput, even when the observable of interest is not itself a dimer free energy.

How these candidate complexes are generated and ranked today, and where that pipeline bottlenecks, is the subject of the next subsection.

A.2. Conformer search and ranking pipelines

Predicting which non-covalent probe–target arrangement dominates requires scanning a high-dimensional pose manifold and ranking candidates whose vacuum electronic binding scores frequently differ by less than 1–2 kcal mol^{-1}. Existing pipelines divide into three families, and each has to pay for its ranking fidelity in a different currency.

Physics-based search on approximate potential-energy surfaces (PESs). The dominant workflow for small, flexible non-covalent interaction (NCI) dimers is CREST (Pracht et al., 2024; Grimme, 2019), widely adopted in the Grimme-group pipeline and community NCI conformer studies. CREST drives RMSD-biased metadynamics and genetic crossover on the semiempirical GFN2-xTB surface (Bannwarth et al., 2019), with an –nci wall restraint that keeps the dimer associated; a GFN-FF pre-screen (Spicher & Grimme, 2020) and a CENSO density functional theory (DFT) re-ranking stage (Grimme et al., 2021a) book-end the search. The broader enhanced-sampling family (replica exchange (Sugita & Okamoto, 1999), metadynamics (Laio & Parrinello, 2002)) shares the same template but typically runs on classical force fields, which systematically misorder halogen-bond, σ-hole, and ionic contacts. CREST is the strong baseline because its search operates on a genuinely quantum-mechanical PES with no ligand/receptor distinction. Its cost is where the bottleneck lives: GFN2-xTB gradients dominate the inner loop, and layering a DFT single-point or re-optimization on top pushes per-system wall-clock into the regime that forecloses high-throughput screening (a claim we make concrete against our own measured CREST+DFT timings in the results), while ab initio molecular dynamics (MD) is ruled out entirely by its per-step DFT cost.

Docking and heuristic pose generation. Docking engines such as AutoDock Vina (Trott & Olson, 2010) and Glide (Friesner et al., 2004) search translational, rotational, and torsional degrees of freedom against a grid-interpolated empirical score tuned on protein–ligand training sets, with deep-learning rescorers (McNutt et al., 2021) and diffusion-based pose samplers (Corso et al., 2023) inheriting the same generate-then-externally-score pattern. They clear 10^3–10^5 poses per CPU-hour, which makes them appropriate for virtual-screening triage, but the returned scores are shape-and-complementarity heuristics calibrated for rigid biomolecular pockets, and the absolute scores are not interpretable as quantitative interaction energies for generic, chemically diverse probe–target dimers with halogens, σ-holes, or charged fragments.

Generative ML pose and conformer models. RDKit's ETKDG (Riniker & Landrum, 2015) produces intramolecular conformers of a single molecule from knowledge-based torsion priors in milliseconds, while GeoDiff (Xu et al., 2022), torsional diffusion (Jing et al., 2022), and NeuralPLexer (Qiao et al., 2024) extend generative diffusion or flow models to dimer and complex geometries at seconds per sample on a single GPU. These models are excellent candidate generators, but they sample from a learned geometric prior and do not emit calibrated electronic binding scores; quantitative ranking still requires an external energy evaluator downstream.

Across all three families the accurate option (CREST-style search on a quantum surface) spends nearly all of its wall-clock inside semiempirical or DFT single-points, while the fast options either return uncalibrated scores or no energies at all. An interatomic-potential surrogate with DFT-level accuracy for the electronic binding score would replace that inner-loop evaluator without forcing a redesign of the outer sampling logic; Section A.3 reviews why pretrained universal MLIPs are beginning to offer the accuracy and coverage needed to test this replacement directly.

A.3. Universal machine learning interatomic potentials as a binding-score surrogate

RAPIDS predicts a pose-ranked relaxed vacuum electronic binding score

$$\Delta E_{\mathrm{bind}} = E_{PT}(R_{PT}^{\mathrm{opt}}) - E_P(R_P^{\mathrm{opt}}) - E_T(R_T^{\mathrm{opt}})$$

on the ωB97M-V/def2-TZVPD surface exposed by the FAIRChem UMA omol head, with an optional ALPB implicit-water correction; this is neither a binding free energy nor a docking score. The question for this subsection is narrow: has the universal machine learning interatomic potential (MLIP) literature (architectures, training data, and pretrained checkpoints) reached the point where using a high-level binding-score evaluator inside an outer pose-search loop is *plausible enough to test empirically*, and what must a screening pipeline do on top of the MLIP to make its outputs trustworthy? The evidence below is proxy evidence (sibling models, covered-chemistry benchmarks, related training sets); whether it carries over to the specific UMA omol head on our 5,565-dimer probe–target regime is the empirical question answered by the benchmarks in the results section of this paper.

From invariant regressors to equivariant message passing. MLIPs evolved from Gaussian kernels and invariant neural networks over hand-crafted descriptors (Behler & Parrinello, 2007; Bartók et al., 2010) to E(3)-equivariant message-passing architectures that carry spherical tensors through every layer. NequIP (Batzner et al., 2022), MACE (Batatia et al., 2022), and Allegro (Musaelian et al., 2023) are representative: by embedding rotational symmetry into the representation rather than learning it, they reach sub-kcal/mol force-field accuracy with roughly an order of magnitude less training data than invariant predecessors. EquiformerV2 later extended this family to higher angular-momentum channels at transformer scale (Liao et al., 2024).

Molecular-chemistry data at training scale. The second enabler was training data that actually covers drug-like, probe–target-relevant chemistry. SPICE provides >1.1 M conformations of small molecules, dimers, dipeptides, and solvated amino acids at ωB97M-D3(BJ)/def2-TZVPPD (Eastman et al., 2023); OMol25 releases over 100 M ωB97M-V/def2-TZVPD single-points across 83 elements with explicit solvation, variable charge, and variable spin, up to 350 atoms per system (Levine et al., 2025); and DES370K supplies CCSD(T)-level dimer interaction energies for out-of-distribution validation (Donchev et al., 2021). Materials-side corpora such as OC20, MPtrj, and OMat24 broadened periodic-table coverage for joint pretraining (Chanussot et al., 2021; Deng et al., 2023; Barroso-Luque et al., 2024) but are peripheral to the dimer interaction energies of interest here.

Universal pretrained models. Several pretrained universal MLIPs now ship as off-the-shelf energy/force evaluators, including MACE-MP-0 for atomistic materials chemistry (Batatia et al., 2024), MACE-OFF23 for organic molecules (Kovács et al., 2025), and UMA, a mixture-of-linear-experts EquiformerV2-style backbone jointly pretrained on OMol25, OC20, OMat24, ODAC23, and OMC25 (Wood et al., 2025). RAPIDS invokes the UMA omol head specifically because it is trained at ωB97M-V/def2-TZVPD with VV10 dispersion, a high-performing NCI DFT level and the OMol25/UMA omol training level, so its native predicted quantity lives at the same electronic-structure level as the relaxed score RAPIDS wants to rank.

What the pretrained model supplies. Four properties matter for inner-loop use. *Accuracy*: MACE-OFF23 reports low errors on neutral closed-shell organic-molecule test sets derived from SPICE and DES370K dimers relative to its training theory (Kovács et al., 2025), and ANI-1ccx approaches CCSD(T)/CBS conformer rankings at neural-network cost (Smith et al., 2019). *Coverage*: UMA and MACE-MP span most of the periodic table with a single checkpoint (Wood et al., 2025; Batatia et al., 2024), and OMol25 extends to charged and open-shell species and to explicit-solvent environments (Levine et al., 2025). *Speed*: for a typical probe–target dimer the dominant cost per energy/force call drops by several orders of magnitude relative to a hybrid-DFT/def2-TZVPD reference on comparable hardware; concrete wall-clock comparisons against CREST+DFT are reported in the results section of this paper. *Differentiability*: analytic forces come from one backward pass, which is what lets RAPIDS drive LBFGS geometry optimization on the MLIP surface without finite differences.

Known failure modes that the pipeline must own. Within covered chemistry a pretrained MLIP can be a promising inner-loop surrogate for DFT single-point evaluation inside a relaxed-pose molecular screening loop when evaluated relative to the same reference functional, but it still comes with important caveats. Three failure modes persist: (i) out-of-distribution extrapolation on unusual oxidation states, rare element combinations, or reactive geometries; (ii) absence of explicit long-range electrostatics in the finite-cutoff local architectures shared by most current backbones, which is consequential for highly polar or charged dimers; and (iii) inherited systematic error of the training functional, so that a UMA omol prediction is at best as accurate as ωB97M-V/def2-TZVPD, and any residual deviation from CCSD(T) on a given motif propagates into every score.

Fine-tuning as an alternative path, and its costs. A natural way to address these failure modes is to fine-tune the foundation MLIP on task-specific data, and a growing literature reports clear gains from this route, including correcting systematic potential-energy-surface softening (Deng et al., 2024), reaching sub-kJ/mol sublimation enthalpies on molecular crystals with limited high-level data (Kaur et al., 2024), and frozen-weight transfer learning for materials chemistry (Radova et al., 2025). Fine-tuning is, however, expensive on two axes that compound. First, it requires curating a downstream-relevant labeled dataset at the target electronic-structure level, which for non-covalent dimers means CCSD(T)/CBS-style or high-level DFT references that are themselves the bottleneck the surrogate was meant to bypass. Second, training and adapting modern equivariant message-passing architectures is itself compute-heavy: EquiformerV2 reports on the order of 1,368 GPU-days for a single OC20 run (Liao et al., 2024), and dedicated systems work has been needed to bring equivariant-MLIP training and inference down to practical cost (Tan et al., 2025). RAPIDS therefore complements rather than replaces this

fine-tuning literature: it asks how far one can push a pretrained universal MLIP at inference time, with no additional training or labeled data beyond a small per-benchmark calibration slice, before fine-tuning becomes worth its data and compute cost.

What RAPIDS adds. A universal pretrained MLIP provides an inner-loop energy surrogate, but it does not by itself perform the pose search, guard against the ML-specific failure modes it introduces, or calibrate against a gold-standard reference. RAPIDS is the scaffold that does. The remainder of the paper lays out the tiered orientation sampling, LBFGS with smart continuation, four per-configuration guards (topology/geometry/energy/convergence) together with a scan-level Energy Consistency Guard (ECG), implicit-solvation post-correction, and benchmark harness that together turn the UMA omol head into a screening-grade probe–target binding-score predictor, benchmarked against CREST+DFT on 5,565 dimers drawn from 18 NCI datasets.

A.4. Multi-fidelity decision-making and offline policy evaluation

A benchmark in which each retained system has been scored by every available method is, by construction, an offline multi-fidelity decision log: each dimer is a context, each fidelity is an arm, and each successfully computed entry of the $5{,}565 \times 9$ result table is a realized (reward, cost) pair. Cheap arms (the RAPIDS MLIP-relaxed score, plain DFT single-points on the MLIP geometry) are fast and sometimes inaccurate; expensive arms (CREST_xTB_DFT) are slow and not uniformly more accurate. As catalogued in §A.2, no single arm dominates the others across the 18 datasets, which motivates treating method selection as an adaptive decision problem rather than a fixed pipeline.

A.4.1. Multi-armed and multi-fidelity bandits

The stochastic multi-armed bandit problem formalizes sequential choice under uncertainty with a logarithmic regret lower bound (Lai & Robbins, 1985), matched by upper-confidence-bound strategies (Auer et al., 2002) and by posterior-sampling strategies whose modern regret analyses (Agrawal & Goyal, 2012) revived an idea of Thompson (1933). Contextual extensions such as LinUCB (Li et al., 2010) condition arm choice on a feature vector, which is the object most similar to the per-system descriptors RAPIDS already computes. Multi-fidelity bandit variants make arm cost and resolution explicit: MF-GP-UCB (Kandasamy et al., 2016) and its continuous-approximation successor BOCA (Kandasamy et al., 2017) use cheap low-fidelity queries to localise the optimum before committing expensive evaluations, and multi-information-source knowledge-gradient acquisition generalizes this to biased surrogates (Huang et al., 2006; Poloczek et al., 2017). The hyperparameter-optimization line (successive halving (Jamieson & Talwalkar, 2016), Hyperband (Li et al., 2018), BOHB (Falkner et al., 2018), and freeze-thaw Bayesian optimization (Swersky et al., 2014)) treats partial training budget as a fidelity axis, and entropy-based acquisition (Hennig & Schuler, 2012) quantifies information gain per unit cost. Multi-fidelity Bayesian optimization has been used for molecular and materials design, where expensive DFT or experimental measurements are paired with cheaper force-field or semi-empirical surrogates (Folch et al., 2023).

A.4.2. Offline policy evaluation and bandit replay

When new online experiments are costly, a logged dataset of (context, action, reward) tuples can be used to estimate the value of counterfactual policies. Importance-sampling estimators reweight logged rewards by the target/logging propensity ratio (Precup et al., 2000); doubly-robust estimators combine importance sampling with a learned reward model to reduce variance while retaining consistency (Dudík et al., 2011; 2014); self-normalized variants and counterfactual risk minimization bound generalization when propensities are small (Swaminathan & Joachims, 2015). For exchangeable logs with a uniform logging policy, Li et al. (2011) showed that a simple "replay" estimator (step through logged events and accept only those whose logged action matches the target policy's choice) is unbiased and has become the standard offline evaluation protocol in recommendation and advertising. These estimators are the machinery a future *online* RAPIDS deployment would need, where only the arm actually selected per system is logged; on the precomputed benchmark used here the factorial (context $\times$ arm) design means they collapse to direct estimation on the subset of cells that completed successfully. The current paper goes one step further and replays deployable rule-based/frozen policies together with retrospective full-information and online-feedback policies, so the method section treats this matrix as an offline replay object first and as a conceptual bandit analogue second.

A.4.3. Why a replay-ready benchmark matters

The $5{,}565 \times 9$ precomputed reward–cost table released with RAPIDS is, operationally, a replay log over probe–target binding-score evaluation (remaining SCF-failure cells are exposed as NaN and handled by a declared fallback in the replay

protocol of §3.4). Any allocation policy that maps an incoming dimer to one fidelity (or to a sequence of fidelities with an early-exit rule) can be evaluated on this log without running a single new DFT job, and comparisons across policies inherit the same benchmark references and wall-clock costs. This subsection provides the vocabulary; the method and results sections instantiate specific policies, from an oracle lower bound through contextual and cost-aware variants, and compare their cost–accuracy curves against that lower bound rather than reporting an explicit cumulative-regret trajectory.

A.5. Large language model agents for autonomous chemistry

Modern large language models have become capable orchestrators of tool-mediated workflows: rather than emitting a final answer in one pass, they decide when to call an external API, how to parse its output, and how to chain results into a longer reasoning trajectory (Schick et al., 2023; Yao et al., 2023b). Chemistry has been one of the earliest natural-science domains where this pattern took hold, because many chemistry sub-problems (structure retrieval, reaction condition lookup, property prediction, simulation submission) already exist as callable tools behind stable interfaces. What the agent layer adds is a way to compose those tools from a natural-language goal.

A.5.1. Autonomous-chemistry agent frameworks

ChemCrow couples GPT-4 to eighteen expert-designed tools spanning literature search, reaction prediction, retrosynthesis, and safety checks, and closes end-to-end loops for target synthesis and chromophore design (Bran et al., 2024). Coscientist pairs an LLM planner with documentation search, Python execution, and a cloud-lab API to carry out palladium-catalyzed cross-coupling optimization on physical hardware (Boiko et al., 2023). Chemist-X targets reaction-condition recommendation with retrieval-augmented generation over molecular databases (Chen et al., 2025), while CACTUS connects open-source LLMs to cheminformatics tools for molecular-property queries and drug-likeness assessment (McNaughton et al., 2024). ChemToolAgent quantifies when tool augmentation helps versus hurts on general chemistry problems (Yu et al., 2025), and ChemAgent introduces a self-updating memory library that improves multi-step chemical reasoning (Tang et al., 2025). Materials-discovery agents follow a similar recipe: ChatMOF couples an LLM to MOF databases and a generator for inverse design (Kang & Kim, 2024); HoneyComb retrieves from a curated materials knowledge base and a tool hub (Zhang et al., 2024); LLMatDesign prompts a reasoning model through iterative materials-property optimization (Jia et al., 2024). Adjacent scientific-reasoning agents such as SciAgent and drug-pipeline agents such as PharmAgents extend the pattern to broader domains (Ma et al., 2024; Gao et al., 2025). Verification strategies from general agent research (self-consistency voting, tree search over reasoning paths, and search-guided fine-tuning) are increasingly imported to reduce tool-call errors and unfounded claims (Wang et al., 2023; Yao et al., 2023a; Putta et al., 2024).

A.5.2. The open gap for non-covalent binding

The frameworks above concentrate on synthesis and retrosynthesis, materials inverse design, or literature-grounded QA. Non-covalent binding, the dominant physics in sensor design, lead optimization, and supramolecular chemistry, remains underserved because every available energy oracle fails one agent-level requirement. A DFT single-point takes minutes to hours per dimer, which is incompatible with the many tool calls a tree-search or self-consistency agent emits per task (LATS, tree-of-thoughts and related orchestrators grow this budget with tree width and sample count); empirical force fields and docking scores are fast but cannot rank the sub-kcal/mol differences that separate competing probes or competing poses for a given target. Commercial AI-assisted discovery platforms expose docking, FEP+, and predictive-model pipelines to chemists, but we are not aware of a published, open-protocol, MLIP-level binding-score oracle callable through an agent interface. RAPIDS addresses that gap: it exposes MLIP-speed binding-score predictions (§A.3), wrapped in validation guards and served through a Model Context Protocol endpoint whose tool surface is described in §A.6 (Anthropic, 2024b). The main contribution here is the oracle, the benchmark matrix, and the agent-ready interface. The paper includes one bounded downstream PFAS harness as an illustration, rather than a large benchmark of agentic orchestration.

A.6. Tool exposure for scientific agents: Model Context Protocol (MCP) and progressive-disclosure skills

Large language models now treat tool use as a first-class capability: a prompt declares external functions, and the model emits structured calls that the host executes and returns to the context window (OpenAI, 2023; Anthropic, 2024c). Raw function-calling scales badly, however: every added tool inflates the prompt, schemas drift across vendors, and invocations are rarely captured in a replayable form. The Model Context Protocol (MCP), together with Anthropic's Agent Skills pattern, which operationalizes *progressive disclosure* for LLM tool exposure, address these failure modes, and together shape how a

scientific tool such as RAPIDS is surfaced to an agent.

A.6.1. From function calling to the Model Context Protocol

Early tool-augmented LLMs interleaved reasoning with external actions (Yao et al., 2023b) and learned to self-select APIs (Schick et al., 2023). Scaling past a handful of tools motivated retrieval-over-APIs benchmarks such as ToolLLM/ToolBench (Qin et al., 2024) and Gorilla (Patil et al., 2023), and orchestration frameworks including LangGraph (LangChain, 2024), AutoGen (Wu et al., 2024), CrewAI (Moura, 2024), and Semantic Kernel (Microsoft, 2024). Tool exposure nonetheless remained ad hoc: each client re-implemented schema plumbing, authentication, and logging. MCP, announced by Anthropic in November 2024 (Anthropic, 2024b; Model Context Protocol Contributors, 2025), standardizes this interface: an MCP *server* advertises three core primitives (tools, resources, and prompts) over JavaScript Object Notation Remote Procedure Call (JSON-RPC), and any MCP *client* can consume them. Current client implementations include VS Code's agent mode (Microsoft, 2025) and the clients catalogued in the MCP specification (Model Context Protocol Contributors, 2025), making a single MCP server reachable from the growing ecosystem of scientific and developer agents.

A.6.2. Progressive disclosure and the Skills pattern

Even with a clean transport, advertising dozens of tools with full schemas saturates the context window. Anthropic's Agent Skills pattern (Zhang et al., 2025; Anthropic, 2025a), described later within the broader Anthropic agent-tooling ecosystem that also includes Computer Use (Anthropic, 2024a), operationalizes *progressive disclosure*: only a short skill name and one-sentence description are loaded up front, while full instructions, schemas, and reference files are fetched on demand. For a scientific package the consequences are concrete. The RAPIDS surface (fourteen MCP tools with numerous parameters each) would otherwise displace task-relevant context, whereas a progressively disclosed summary remains near-free until invoked. Progressive disclosure is also compatible with client-side audit logging: hosts such as Claude Desktop and Claude Code log each tool invocation's typed inputs and outputs, so that together with server-reported versioning a reviewer can reconstruct which tool version produced which number, a deployment property the MCP transport makes straightforward but does not itself guarantee.

A.6.3. Consequences for RAPIDS

MLIP-based binding prediction is only useful to an agent if it is callable with a clean interface, a small prompt footprint, and logs that downstream clients can audit. MCP provides the transport; progressive disclosure, whether via a full Skills wrapper or via MCP tools whose surface area is deliberately minimized, keeps the prompt footprint small. RAPIDS packages its guarded pose search, validation guards, fidelity catalog, and offline-replay artifacts as a small Skill-level surface over the underlying MCP server, so an agent can call the live oracle, inspect the nine-fidelity ladder, and request replay-based allocation advice through the same interface. Agent-side orchestration is delegated to the frameworks and clients surveyed in §A.5, and the concrete tool inventory is documented in §3.2.

B. Detailed Methods

This appendix records the exact quantity definitions, initialization choices, thresholds, pseudocode, and protocol details that support the compact main-text Methods section. Both pipelines evaluated in this paper target the same predicted quantity: a relaxed vacuum electronic binding energy

$$\Delta E_{\mathrm{bind}} = E_{PT}(R_{PT}^{\mathrm{opt}}) - E_P(R_P^{\mathrm{opt}}) - E_T(R_T^{\mathrm{opt}})$$

for a probe–target dimer on the ωB97M-V/def2-TZVPD (Mardirossian & Head-Gordon, 2016; Weigend & Ahlrichs, 2005; Rappoport & Furche, 2010) high-level functional surface. Because both RAPIDS and CREST+DFT relax the complex and the isolated monomers before subtraction, this is *not* a strict frozen-geometry interaction energy. Throughout the paper the benchmark target remains the published interaction-energy reference

$$\Delta E_{\mathrm{int}}^{\star} = E_{PT}(R_{PT}^{\star}) - E_P(R_P^{PT,\star}) - E_T(R_T^{PT,\star}),$$

so every reported MAE should be read as a benchmark-reference deviation after method-specific pose relaxation. The appendix sections below expand the compact main-text summaries in §3.2–§3.4.

Table 1. Benchmark map for the 5,565 completed systems used in the paper. Counts are completed-system counts. The last column records the published benchmark level treated as ground truth in this work together with the original dataset citation.

Benchmark	N	Charge	Representative chemistry	Published ground truth
A24	22	Neutral	Canonical small-molecule dimers.	CCSD(T)/CBS-style interaction energies; A24 benchmark (Řezáč et al., 2015).
S66	66	Neutral	Balanced H-bond, dispersion, and mixed organic dimers.	CCSD(T)/CBS interaction energies; S66 benchmark (Rezac et al., 2011).
X40	40	Neutral	Halogen-bond dimers.	CCSD(T)/CBS interaction energies; X40 benchmark (Rezac et al., 2012).
HB300SPX	300	Neutral	Extended hydrogen-bond set with heavier donor and acceptor chemistry.	NCIAtlas CCSD(T)/CBS-style interaction energies; HB300SPX (Řezáč, 2020b).
HB375	375	Neutral	Diverse neutral hydrogen-bond dimers.	NCIAtlas CCSD(T)/CBS-style interaction energies; HB375 (Řezáč, 2020a).
SH250	250	Neutral	σ-hole interactions across halogen, chalcogen, and pnictogen donors.	NCIAtlas CCSD(T)/CBS-style interaction energies; SH250 (Kříž & Řezáč, 2022).
D1200_HBCNO	308	Neutral	London-dispersion subset containing only H, B, C, N, and O.	NCIAtlas CCSD(T)/CBS-style interaction energies; D1200 (Řezáč, 2022).
D1200_Halogens	308	Neutral	Halogen-containing London-dispersion subset.	NCIAtlas CCSD(T)/CBS-style interaction energies; D1200 (Řezáč, 2022).
D1200_PS	293	Neutral	P/S-containing London-dispersion subset.	NCIAtlas CCSD(T)/CBS-style interaction energies; D1200 (Řezáč, 2022).
BFDb_BBI	100	Neutral	Protein backbone–backbone fragments.	BFDb CCSD(T)-F12 silver-standard interaction energies (Burns et al., 2017).
BFDb_HSG	21	Neutral	Host–substrate–glue protein fragments.	BFDb CCSD(T)-F12 silver-standard interaction energies (Burns et al., 2017).
BFDb_NBC1	10	Neutral	Nucleobase stacking fragments.	BFDb CCSD(T)-F12 silver-standard interaction energies (Burns et al., 2017).
BFDb_SSI_dispersion	1325	Neutral	Sidechain pairs dominated by dispersion.	BFDb CCSD(T)-F12 silver-standard interaction energies (Burns et al., 2017).
BFDb_SSI_mixed	765	Neutral	Sidechain pairs with mixed polar and dispersion character.	BFDb CCSD(T)-F12 silver-standard interaction energies (Burns et al., 2017).
BFDb_SSI_other	240	Neutral	Sidechain pairs outside the polar, mixed, and dispersion buckets.	BFDb CCSD(T)-F12 silver-standard interaction energies (Burns et al., 2017).
BFDb_SSI_polar	266	Neutral	Polar and electrostatic sidechain pairs.	BFDb CCSD(T)-F12 silver-standard interaction energies (Burns et al., 2017).
IHB100	100	Charged	Ionic hydrogen bonds with formally charged fragments.	NCIAtlas CCSD(T)/CBS-style interaction energies; IHB100 set in the HB/IHB benchmark family (Řezáč, 2020a).
BFDb_SSI_charged	776	Charged	Salt bridges, charged sidechains, and same-charge repulsion pairs.	BFDb CCSD(T)-F12 silver-standard interaction energies (Burns et al., 2017).

For benchmark evaluation on the 5,565-system completed split[1], PubChem retrieval (Kim et al., 2023) is bypassed: monomer coordinates, charges, and the benchmark reference geometry are taken from the dataset-provided dimer records, the two monomers are separated from that record, and each pipeline then reinitializes the complex from those monomers according to its own search rule. For a fixed system and fidelity, the isolated probe and target monomers are evaluated or optimized once with that method's charge/spin assignment and reused when scoring candidate poses of the same system.

B.1. Detailed RAPIDS Implementation

RAPIDS (Rapid Atomistic Probe–target Interaction Discovery Scaffold) builds on the externally pretrained UMA omol head (Wood et al., 2025; Levine et al., 2025). The work here contributes the orientation sampler, the optimizer with smart continuation, the validation stack, the solvation post-correction, the offline-replay harness, and the application-layer MCP/Skills packaging described below. To the best of our knowledge, this is the first systematic use of a pretrained universal MLIP as an interaction-system scaffold benchmarked at this scale and carried into downstream probe-discovery workflows. The implementation supports probe–target–substrate assemblies, but the benchmarked pathway in this paper is the vacuum probe–target dimer case.

Execution scope. The benchmarked RAPIDS path is the guarded screening workflow: structure building, tiered search, LBFGS relaxation, per-configuration validation, scan-level ECG aggregation, and optional solvation.

Energy surface. Every energy or force call is a forward (forward–backward for gradients) pass through the UMA omol head, a mixture-of-linear-experts EquiformerV2-style backbone pretrained at ωB97M-V/def2-TZVPD with VV10 dispersion on OMol25 (Wood et al., 2025; Levine et al., 2025). The head is used here only as a surrogate energy/force evaluator;

[1]The underlying collection contains four additional unsupported cases, two A24 systems and two D1200_Halogens CI_4 systems, for which at least one lower-level engine in the pipeline stack does not support the chemistry.

it does not run explicit DFT at inference time. No self-consistent-field (SCF) iterations, no basis-set selection, and no finite-difference gradients are performed at inference time.

Tiered orientation sampling. The pose search is a three-tier structured scan rather than a stochastic metadynamics. Tier 1 always runs and enumerates $n_{\text{anchors}} \times n_{\text{orientations}} = 3 \times 3 = 9$ probe placements using three anchor offsets and three coarse orientations. Tier 2 adds six basin-hopping perturbations of the best Tier-1 configuration, with randomized rotations and translations around that candidate. Tier 3 adds nine random refinements near the running best. Tier 1→2 upgrades fire if any of four criteria hold on the eligible Tier-1 set: $\sigma_E > 0.3$ eV (≈ 6.9 kcal/mol), top-two gap < 0.02 eV (≈ 0.46 kcal/mol), eligible fraction $< 80\%$, or system size above 50 atoms; Tier 2→3 fires if Tier 2 improves the Tier-1 minimum by more than 0.1 eV (≈ 2.3 kcal/mol). Caller-forced minimum tiers ($\text{min_tier} \geq 2$ or ≥ 3) override these triggers, whereas an empty eligible Tier-1 set blocks any upgrade.

Formally, letting $\mathcal{R}_1$ denote the Tier-1 result set, define

$$\mathcal{R}_1^{\text{elig}} = \{r \in \mathcal{R}_1 : c(r) \in \{\text{high}, \text{medium}\}\}, \tag{1}$$

$$\sigma_1 = \sigma\Big(\{E(r) : r \in \mathcal{R}_1^{\text{elig}}\}\Big), \tag{2}$$

$$\eta_1 = \frac{|\mathcal{R}_1^{\text{elig}}|}{|\mathcal{R}_1|}, \tag{3}$$

$$\Delta E_{1,2} = E_2^{(1)} - E_{\min}^{(1)}. \tag{4}$$

We use the conventions $\sigma_1 = 0$ if $|\mathcal{R}_1^{\text{elig}}| = 0$, $\Delta E_{1,2} = +\infty$ if $|\mathcal{R}_1^{\text{elig}}| < 2$, $\eta_1 = 0$ if $|\mathcal{R}_1| = 0$, and $E_{\min}^{(2)} = +\infty$ if the eligible Tier-2 set $\mathcal{R}_2^{\text{elig}}$ is empty. The two upgrade gates are then

$$U_{1\to2} = \begin{cases} 1, & \text{if } \max_\text{tier} \geq 2,\ |\mathcal{R}_1^{\text{elig}}| > 0, \text{ and at least one of the following holds: } \sigma_1 > 0.3 \text{ eV;} \\ & \Delta E_{1,2} < 0.02 \text{ eV; } \eta_1 < 0.8;\ N_{\text{atoms}} > 50;\ \text{min_tier} \geq 2, \\ 0, & \text{otherwise,} \end{cases}$$

$$U_{2\to3} = \begin{cases} 1, & \text{if } \max_\text{tier} \geq 3 \text{ and either } \text{min_tier} \geq 3 \text{ or } \big(|\mathcal{R}_2^{\text{elig}}| > 0 \wedge E_{\min}^{(2)} < E_{\min}^{(1)} - 0.1 \text{ eV}\big), \\ 0, & \text{otherwise.} \end{cases}$$

The second gate uses the best eligible Tier-2 energy only when such a candidate exists. These definitions are implementation-faithful summaries of the prose trigger rules above.

Local optimization. Each configuration is relaxed with limited-memory BFGS on the MLIP surface to $f_{\max} = 0.05$ eV Å^{-1}, capped at 100 steps per run with up to three *smart continuations* that re-enter optimization from the best previously seen geometry.

Validation guards. RAPIDS separates per-configuration guards from scan-level aggregation. The per-configuration guard stack contains topology, geometry, energy, and convergence checks. The *topology guard* compares covalent bond graphs of the initial and final structures (ase.neighborlist (Larsen et al., 2017) with element-pair thresholds and a 15% bond-change tolerance). If an intramolecular bond change is detected on a neutral system, the worker re-optimizes the same geometry with fallback UMA heads and downgrades confidence by majority vote; for systems with any nonzero fragment or complex charge ($q_P \neq 0$, $q_T \neq 0$, or $q_{PT} \neq 0$), the worker downgrades confidence directly and returns early without executing the later guards. When topology is preserved, the *geometry guard* flags severe non-bonded contacts via the MolProbity criterion (Chen et al., 2010) $d_{ij} < r_i^{\text{vdW}} + r_j^{\text{vdW}} - 0.4$ Å together with strained intramolecular bonds. The *energy guard* writes $E_{\text{bind}} \equiv \Delta E_{\text{bind}}$ and applies two thresholds using the smaller fragment atom count $N_{\text{small}} = \min(|P|, |T|)$ rather than the full complex size: $|E_{\text{bind}}|/N_{\text{small}} \leq 0.65$ eV/atom and $|E_{\text{bind}}| \leq 2.2$ eV (≈ 50.7 kcal/mol). The benchmarked flow uses this two-threshold rule. The absolute 2.2 eV ceiling is an empirical catastrophic-failure threshold aligned with the 50 kcal/mol clipping scale used in evaluation. That clipping scale lies well above even very strong monovalent host–guest binding free energies, which are on the order of $\sim$ 20–24 kcal/mol for benchmark CB[7] complexes (Cao et al., 2014; Shetty et al., 2015), and it is also consistent with chemistry-ML practice that treats catastrophic thresholds pragmatically and regards energy deviations of roughly 40–56 kcal/mol as unreasonably large or pathological (Bhattacharjee et al., 2021;

Chung & Green, 2024). Finally, the *convergence guard* downgrades only topology-preserved runs, assigning low to diverged smart-continuation sequences and capping step-limit exits at medium. These guard thresholds live in the worker's native eV units and act as catastrophic-failure sentinels rather than as ranking thresholds in the 1–2 kcal/mol regime discussed in §A.1.

At the level of abstraction used in the paper, let $b_{\text{early}}(r) = 1$ denote the broken-topology branch that returns before the later guards, and $b_{\text{early}}(r) = 0$ otherwise. Each candidate r then determines the executed-guard index set

$$\mathcal{G}_{\text{exec}}(r) = \begin{cases} \{\text{topo}\}, & b_{\text{early}}(r) = 1, \\ \{\text{topo}, \text{geom}, \text{en}, \text{conv}\}, & b_{\text{early}}(r) = 0. \end{cases}$$

The emitted alert set is then

$$\mathcal{A}(r) = \{\, g_k(r) : k \in \mathcal{G}_{\text{exec}}(r) \,\}.$$

The energy alert itself is

$$\begin{aligned} g_{\text{en}}(r) &= \mathbf{1}\left[\frac{|E_{\text{bind}}(r)|}{N_{\text{small}}(r)} > 0.65 \text{ eV/atom} \vee |E_{\text{bind}}(r)| > 2.2 \text{ eV}\right], \\ E_{\text{bind}}(r) &\equiv \Delta E_{\text{bind}}(r), \\ N_{\text{small}}(r) &= \min(|P|, |T|). \end{aligned}$$

The reported confidence is an order-sensitive mapping

$$c(r) = \phi\big(\mathcal{A}(r)\big),$$

where unrecoverable topology breaks and non-physical energies force low, while clash-only or step-limit events cap confidence at medium. In the broken-topology branch, the worker returns immediately after fallback-task consensus and the later guards are not evaluated.

Energy Consistency Guard. The scan-level *Energy Consistency Guard* (ECG) is executed only in the MCP scan_orientations path, not in standalone CLI runs. It operates on the high/medium-confidence candidate set

$$\mathcal{R}^{\text{elig}} = \{r : c(r) \in \{\text{high}, \text{medium}\}\},$$

using the geometric-similarity predicate

$$S(r_i, r_j) = \mathbf{1}\left[\Delta\text{COM}(r_i, r_j) \le 1.5 \text{ Å} \;\wedge\; \Delta d_{\min}(r_i, r_j) \le 0.5 \text{ Å}\right],$$

and the similarity-restricted energy span

$$\Delta_{\text{span}} = \max_{i,j:\, S(r_i,r_j)=1} |E_{\text{bind}}(r_i) - E_{\text{bind}}(r_j)|,$$

with the convention $\Delta_{\text{span}} = 0$ when no pair satisfies the similarity predicate, and

$$E_{\min} = \min_{r \in \mathcal{R}^{\text{elig}}} E_{\text{bind}}(r).$$

Let

$$b_{\text{sim}} = \mathbf{1}[\exists\, i, j :\; S(r_i, r_j) = 1].$$

The ECG mode is then

$$m_{\text{ECG}} = \begin{cases} \text{manual_review}, & \Delta_{\text{span}} > 10 \text{ eV} \;\wedge\; b_{\text{sim}} = 1, \\ \text{median}, & \Delta_{\text{span}} > 3 \text{ eV} \;\wedge\; b_{\text{sim}} = 1, \\ \text{median}, & E_{\min} < Q_1 - 3\,\text{IQR}, \\ \text{minimum}, & \text{otherwise}. \end{cases}$$

with committed energy

$$E^{\star} = \begin{cases} \min\limits_{r \in \mathcal{R}^{\text{elig}}} E_{\text{bind}}(r), & m_{\text{ECG}} = \text{minimum}, \\ \text{median}_{r \in \mathcal{R}^{\text{elig}}}\, E_{\text{bind}}(r), & m_{\text{ECG}} = \text{median}. \end{cases}$$

When $m_{\text{ECG}} = \text{manual_review}$, the run is flagged for DFT follow-up and the worker reports both the minimum and median summaries rather than forcing a single committed scalar. The ECG thresholds (3 and 10 eV, i.e. ≈ 69 and 231 kcal/mol) are deliberately far above the ranking scale of interest and serve only as catastrophic-instability sentinels.

Algorithm 1 RAPIDS guarded scan

Retrieve molecules and build the initial probe–target(-substrate) complex.
Run Tier 1 grid search ($3 \times 3 = 9$ placements), optimizing and validating each candidate with Algorithm 2.
if the eligible Tier 1 set is empty **then**
  Return the current result set without upgrading.
end if
if $\sigma_E > 0.3$ eV **or** top-two gap < 0.02 eV **or** valid fraction $< 80\%$ **or** $N_{\text{atoms}} > 50$ **or** min_tier≥ 2 **then**
  Run Tier 2 basin hopping (6 perturbed configurations around the Tier 1 best).
end if
if Tier 2 improves the Tier 1 minimum by more than 0.1 eV **or** min_tier≥ 3 **then**
  Run Tier 3 random refinement.
end if
Apply the ECG to the high/medium-confidence set and choose the final scan selection mode.
if vacuum mode with solvation requested **then**
  Attach the xTB+ALPB correction on the optimized probe, target, and complex geometries.
end if
Return the ranked result set, committed energy, confidence annotations, and run artifacts.

Algorithm 2 Worker-side optimize and validate

Relax the candidate on the UMA omol surface with LBFGS and up to three smart continuations.
Initialize confidence to high and the alert set to empty.
Run the topology guard on the initial and final bond graphs.
if topology is not preserved **then**
  if the system is neutral and fallback UMA heads are available **then**
    Re-optimize with the fallback heads and downgrade confidence by the majority verdict.
  else
    Downgrade confidence directly and mark the run as DFT-recommended.
  end if
  Return immediately; later guards are skipped in the broken-topology branch.
end if
Run the geometry guard and flag severe clashes or strained bonds.
Run the energy guard using $|E_{\text{bind}}|/N_{\text{small}} \leq 0.65$ eV/atom and $|E_{\text{bind}}| \leq 2.2$ eV.
Apply convergence downgrades only on topology-preserved runs.
Return the optimized geometry, predicted energy, confidence, and alert record.

Implicit solvation. A vacuum-mode run can additionally attach a single-point GFN2-xTB (Bannwarth et al., 2019) calculation with the ALPB implicit-water model (Ehlert et al., 2021) on the optimized probe, target, and complex geometries. The solvation correction is

$$\Delta\Delta G_{\text{solv}} = \left(G_{PT}^{\text{aq}} - G_{PT}^{\text{gas}}\right) - \left(G_{P}^{\text{aq}} - G_{P}^{\text{gas}}\right) - \left(G_{T}^{\text{aq}} - G_{T}^{\text{gas}}\right),$$

and the reported ALPB-solvation-corrected electronic score is

$$\Delta G_{\text{sol}} = \Delta E_{\text{bind}}^{\text{vac}} + \Delta\Delta G_{\text{solv}}.$$

We write the three-body correction as $\Delta\Delta G_{\text{solv}}$ to distinguish it from the final score ΔG_{sol}. This quantity omits entropy, thermal, and standard-state corrections and should not be read as a full solution-phase binding free energy. The benchmark metrics in this paper remain vacuum electronic binding scores unless solvation is explicitly requested.

Output. Each simulation records the optimized geometries, trajectories, binding energies, solvation corrections when requested, convergence status, per-guard validation outcomes, and a confidence-annotated best-pose ΔE_{bind}.

Algorithmic summary. Algorithms 1 and 2 summarize the guarded MCP execution path used throughout the paper.

B.2. Detailed CREST + Composite-DFT Baseline Settings

The baseline is a cost-controlled CREST + composite-DFT high-fidelity pipeline inspired by the three-stage recipe of Bursch et al. (Bursch et al., 2022): semiempirical conformer sampling, tight semiempirical reoptimization, and a composite-DFT two-step treatment.

Stage 1: CREST NCI sampling. Initial probe–target dimers are built by translating the monomers along the z axis so that their centres of mass are separated by 4.0 Å. CREST (Pracht et al., 2024) then runs GFN2-xTB (Bannwarth et al., 2019) metadynamics inside a soft logfermi wall that confines the two partners, giving conformers within a 6 kcal mol^{-1} window above the putative global minimum without imposing a reference topology. Duplicates are removed internally by energy and RMSD. If CREST produces no conformers (e.g. convergence failure), the 4.0 Å starting geometry is forwarded to Stage 2 as a fallback.

Stage 2: GFN2-xTB tight reoptimization. The ten lowest-energy CREST conformers and each isolated monomer are reoptimized with tight GFN2-xTB settings. The lowest-energy xTB-optimized complex is retained as the Stage-3 input, and an xTB-level binding energy ΔE_{xTB} is recorded as an intermediate.

Stage 3: GPU4PySCF composite DFT. The retained probe, target, and complex geometries are each treated with a two-step GPU4PySCF workflow (Sun et al., 2020; Li et al., 2025). The first step reconstructs the composite r^{2}SCAN-3c rung with the modified def2-mTZVP basis together with D4 and gCP corrections (Grimme et al., 2021b; Caldeweyher et al., 2019; Kruse & Grimme, 2012). The second step evaluates the final single-point energy at the reference ωB97M-V/def2-TZVPD level in GPU4PySCF. For hybrid and meta-GGA evaluations we use RI-J together with seminumerical exchange, which plays the same acceleration role as RIJCOSX-style approximations without sharing the same implementation or keyword surface. The parsed species energies are combined as a relaxed binding energy $\Delta E_{\mathrm{bind}}^{\mathrm{DFT}} = E_{PT} - E_P - E_T$.

Intermediate DFT arms used in the 9-arm matrix. The six DFT intermediate arms all start from the RAPIDS-committed complex geometry and are all executed in GPU4PySCF. The three single-point (SP) arms evaluate the RAPIDS-optimized complex together with the corresponding RAPIDS-optimized isolated monomers at one explicit single-point level: PBE-D3BJ/def2-TZVP SP, ωB97X-D3BJ/def2-TZVP SP, and ωB97M-V/def2-TZVP SP (Perdew et al., 1996; Grimme et al., 2010; 2011; Chai & Head-Gordon, 2008; Mardirossian & Head-Gordon, 2016). The three geometry-reoptimization plus single-point (GeoSP) arms first reoptimize the complex and the isolated probe and target monomers once at the same functional with def2-TZVP and TightOpt/TightSCF, then evaluate a same-functional def2-TZVPD single-point on that optimized geometry: PBE-D3BJ GeoSP, ωB97X-D3BJ GeoSP, and ωB97M-V GeoSP. In all six arms the fragment charges q_P, q_T, and q_{PT} are carried explicitly, closed-shell singlets are assumed unless the dataset requires otherwise, and no counterpoise/BSSE correction is applied.

Charge, spin, and resources. Complex and monomer charges are set independently: the workflow carries q_P, q_T, and q_{PT} separately, so a salt bridge can have $q_{PT} = 0$ but charged fragments. Unless otherwise noted systems are closed-shell singlets. No symmetry, thermal, or solvation corrections are applied; reported baseline energies are relaxed vacuum electronic binding energies rather than strict frozen-geometry interaction energies. RAPIDS inference and every explicit DFT evaluation in this paper are run locally, while CREST and xTB remain CPU-side preprocessing. All reported costs are wall-clock times for this implementation and are not normalized to energy draw or cloud-dollar cost.

B.3. Detailed Pipeline Correspondence

Table 2 aligns the two pipelines stage by stage.

The two pipelines share two points of agreement and diverge on four axes. Both target the same high-level functional rung, ωB97M-V/def2-TZVPD, and both restrict themselves to vacuum electronic energies unless solvation is explicitly requested. They differ on (i) the evaluator at inference (UMA omol surrogate inference against explicit SCF plus geometry optimization), which drives wall-clock cost; (ii) the sampler (a structured, cached, upgrade-triggered tiered scan against metadynamics at a fixed energy window), which trades exploration for reproducibility; (iii) the committed geometry (a minimum on the UMA omol surface against a minimum on the r^{2}SCAN-3c surface), so that an observed score gap on common systems combines two distinct contributions, the MLIP's fixed-geometry surrogate error against the explicit SP rung used in the matrix and the geometric deviation between UMA-optimized and r^{2}SCAN-3c-optimized complexes; and (iv) the failure-mode model, namely ML-specific guards with multi-head consensus against SCF and optimization convergence diagnostics. The published benchmark target against which errors are reported is neither pipeline: it is the dataset-provided interaction-energy reference at the benchmark geometry. The results section therefore reports benchmark-reference deviations of method-specific relaxed binding scores, and uses intermediate fidelities that apply the reference DFT (as a single-point or as a geometry-reoptimization plus single-point) on top of RAPIDS-committed geometries to separate the surface and geometry contributions. Section 3.4 additionally describes how the full $5{,}565 \times 9$ precomputed result matrix

Table 2. Stage-by-stage correspondence between RAPIDS and the CREST + DFT high-fidelity baseline pipeline. Both produce relaxed vacuum electronic binding energies, but RAPIDS does so with a UMA surrogate trained at the ωB97M-V/def2-TZVPD level while the baseline evaluates explicit DFT through GPU4PySCF. The gold reference used in the results is instead the published benchmark interaction energy at the benchmark geometry.

Stage	**RAPIDS**	**CREST + DFT**
Evaluator used in search	UMA omol (single MLIP pass)	GFN2-xTB (Stages 1–2), GPU4PySCF DFT (Stage 3)
Pose search	Tiered $9 + 6 + 9$ scan with basin-hopping and trigger/min-tier upgrades	GFN2-xTB metadynamics, 6 kcal/mol window
Geometry relax	LBFGS on MLIP, smart continuation	xTB tight $\rightarrow$ assembled r^{2}SCAN-3c TightOpt
Reported electronic score	UMA omol surrogate trained at ωB97M-V/def2-TZVPD	Explicit ωB97M-V/def2-TZVPD via GPU4PySCF
Solvation	Optional xTB+ALPB post-correction in vacuum mode	Vacuum only
Failure checks	Per-config guards + ECG (MCP path)	SCF + optimization convergence

is treated as an offline multi-fidelity replay log for comparing allocation heuristics and replay policies.

B.4. Telescoping decomposition: full per-system definitions

The compact decomposition stated in §4.3 expands as follows. For each system s, the four scores are

$$\hat{y}_{\mathrm{R}}(s) = \hat{\Delta E}_{\mathrm{bind}}^{\mathrm{RAPIDS}}(s), \tag{5}$$

$$\hat{y}_{\mathrm{SP}}(s) = \hat{\Delta E}_{\mathrm{bind}}^{\mathrm{RAPIDS_wB97M-V_SP}}(s), \tag{6}$$

$$\hat{y}_{\mathrm{Geo}}(s) = \hat{\Delta E}_{\mathrm{bind}}^{\mathrm{RAPIDS_wB97M-V_GeoSP}}(s), \tag{7}$$

$$\hat{y}_{\mathrm{C}}(s) = \hat{\Delta E}_{\mathrm{bind}}^{\mathrm{CREST_xTB_DFT}}(s). \tag{8}$$

The exact telescoping identity at the system level is

$$\hat{y}_{\mathrm{R}}(s) - \hat{y}_{\mathrm{C}}(s) = \delta_{\mathrm{surf}}(s) + \delta_{\mathrm{geom}}(s) + \delta_{\mathrm{samp}}(s),$$

with the three axis-aligned increments

$$\delta_{\mathrm{surf}}(s) = \hat{y}_{\mathrm{R}}(s) - \hat{y}_{\mathrm{SP}}(s), \tag{9}$$

$$\delta_{\mathrm{geom}}(s) = \hat{y}_{\mathrm{SP}}(s) - \hat{y}_{\mathrm{Geo}}(s), \tag{10}$$

$$\delta_{\mathrm{samp}}(s) = \hat{y}_{\mathrm{Geo}}(s) - \hat{y}_{\mathrm{C}}(s). \tag{11}$$

B.5. Detailed Offline-Replay Protocol

Purpose. Sections 3.2–3.3 fix two full pipelines at one operating point each. Here we replay the resulting $5{,}565 \times 9$ prediction/time matrix as an offline multi-fidelity decision log, so arbitrary allocation policies over the same nine arms are scored on the same benchmark stream without issuing new DFT jobs (forward reference §A.4).

Setup. A context s_t is one of 5,565 probe–target systems, drawn in permutation π_k (seed k) from the 18 benchmarks. The arm set $\mathcal{A}$ is the nine fidelities of §3.3: RAPIDS, three DFT_SP on RAPIDS geometry, three DFT_GeoSP, CREST_xTB, CREST_xTB_DFT. Each log cell holds a predicted relaxed binding score $\hat{\Delta E}_{\mathrm{bind}}(s_t, a)$, wall-clock time $\tau(s_t, a)$, and a status flag; $\Delta E^{\star}_{\mathrm{int}}(s_t)$ is the published benchmark interaction-energy reference (CCSD(T)/CBS or equivalent). Reward is capped negative absolute benchmark-reference deviation,

$$r(s_t, a) = -\min\big(|\hat{\Delta E}_{\mathrm{bind}}(s_t, a) - \Delta E^{\star}_{\mathrm{int}}(s_t)|,\ E_{\max}\big),$$

with $E_{\max} = 50$ kcal/mol, and cost is $c(s_t, a) = \tau(s_t, a)$ in seconds. Reference values are never exposed at arm-selection time, but some replayed policy families do consume them during retrospective updates as described below.

Policy formalization. At each step t, a replay policy π maps the current context to an ordered arm sequence,

$$\mathcal{M}_t = \pi(s_t) \in \mathcal{A}^*,$$

where $\mathcal{A}^*$ denotes the set of all finite ordered sequences over the arm set $\mathcal{A}$. Writing

$$\mathcal{M}_t = (a_{t,1}, \ldots, a_{t,m_t}),$$

with committed arm

$$a_t^\dagger = \text{last_nonNaN}(\mathcal{M}_t),$$

committed prediction

$$\hat{y}_t^\pi = \hat{\Delta E}_{\text{bind}}(s_t, a_t^\dagger),$$

and paid cost

$$c_t^\pi = \sum_{j=1}^{m_t} \tau(s_t, a_{t,j}).$$

The replay reward is

$$r_t^\pi = -\min\big(|\hat{y}_t^\pi - \Delta E_{\text{int}}^\star(s_t)|,\, E_{\max}\big),$$
$$E_{\max} = 50 \text{ kcal/mol},$$

and the corresponding capped error objective is

$$\text{MAE}_{\text{cap}}(\pi) = \frac{1}{T} \sum_{t=1}^{T} \min\big(|\hat{y}_t^\pi - \Delta E_{\text{int}}^\star(s_t)|,\, E_{\max}\big).$$

We therefore interpret replay as a budgeted multi-fidelity allocation problem,

$$\min_\pi \quad \text{MAE}_{\text{cap}}(\pi)$$
$$\text{s.t.} \quad \mathbb{E}\left[\sum_{t=1}^{T} c_t^\pi\right] \le B,$$

while reporting the empirical Pareto frontier across policies rather than committing to a single budget B. This optimization view is the underlying decision problem that replay approximates; the paper's reported outputs remain empirical cost–MAE/ρ comparisons among concrete policies.

Policies replayed. Four families are implemented. *Static*: nine ALWAYS-X policies and an ORACLE lower bound that, for each s_t, chooses $\arg\min_a |\hat{\Delta E}_{\text{bind}} - \Delta E_{\text{int}}^\star|$ while paying only that arm's cost. *Context-free selectors*: RANDOM; UCB1 (Auer et al., 2002) on (RAPIDS-magnitude bin, arm) pairs over five quantile bins on $|\hat{\Delta E}_{\text{bind}}^{\text{RAPIDS}}|$, with and without a cost-normalized score (UCB statistic divided by each method's median cost); Thompson sampling (Thompson, 1933) on the same bins; an explore-then-exploit COST-AWARE policy that pulls all nine arms on the first 10% of systems and then greedily selects arms according to predicted benefit per unit cost; and MFMI-GREEDY. *Escalation / stacking*: DISAGREEMENT, PROGRESSIVE LADDER (RAPIDS $\rightarrow$ PBE-D3BJ_SP $\rightarrow$ ωB97M-V_SP $\rightarrow$ ωB97M-V_GeoSP $\rightarrow$ CREST_xTB_DFT), online STACKING, STACKING ML, CHEAP ENSEMBLE, BIAS CORRECTION. *Context-aware / meta*: LEARNED SELECTOR (trained on RAPIDS-value features only), CHEMUCB on eight chem-feature bins, CHEM LEARNED SELECTOR (the chem-feature twin of LEARNED SELECTOR), ALORS/CHEMALORS (Mişir & Sebag, 2017), LINUCB (Li et al., 2010) on chemical plus cheap-prediction features, leave-one-benchmark-out BUCKET PRIOR (adaptive and BUCKET LEARNED variants), and a gradient-boosted-machine (Friedman, 2001) family (GBM STACKING, META-GBM, META-GBM ADAPTIVE, GBM META-WARM, HYBRID SELECT+STACK). Priors and meta-models are trained leave-one-benchmark-out. The reported study compares 40 replay policies spanning these families.

Per-system selection. Single-shot and escalating policies share one protocol: on each s_t a policy returns an ordered sequence $\mathcal{M}(s_t) \in \mathcal{A}^*$; the committed prediction is the last non-NaN value in $\mathcal{M}(s_t)$ and the paid cost is $\sum_{j=1}^{m_t} \tau(s_t, a_{t,j})$ for $\mathcal{M}(s_t) = (a_{t,1}, \ldots, a_{t,m_t})$. Static and context-free selectors use $m_t = 1$ after any warm-up phase. Escalation policies extend $\mathcal{M}$ only when an explicit disagreement trigger fires; PROGRESSIVE LADDER, for example, applies three sequential disagreement thresholds on consecutive-prediction gaps at successive ladder stages and sweeps four threshold triples. Stacking policies instead run a fixed cheap-arm prefix on every system: STACKING always executes a short list of cheap arms and only then escalates on disagreement, while the STACKING ML / GBM STACKING / META-GBM meta-learners are non-triggered: they always call the same cheap-feature prefix and fuse its outputs through a learned combiner, paying the full prefix cost on every s_t. For the full-information selector families, posterior or weight updates use the all-arm error vector available in the precomputed matrix; this is retrospective full-information replay, not selected-arm bandit feedback.

Feedback regimes and deployability. The replay suite mixes four feedback assumptions. Rule-based policies (ALWAYS-X, DISAGREEMENT, PROGRESSIVE LADDER, CHEAP ENSEMBLE) never consume benchmark references during replay-time updates and are directly deployable as written. META-GBM is a frozen leave-one-benchmark-out meta-model and is also directly deployable. Online regression / bias-correction policies (STACKING, STACKING ML, GBM STACKING, BIAS CORRECTION, META-GBM ADAPTIVE) update after each processed system using that system's benchmark reference and therefore require a delayed ground-truth feedback loop to deploy. UCB/Thompson/ChemUCB/BucketPrior/Learned-Selector-style policies receive full-information retrospective updates from the errors of all nine arms on the current system because the factorial log contains every arm; they should be read as retrospective selectors rather than as causal online bandits. We keep all four classes because together they map the achievable cost–accuracy frontier on the same matrix, but only the rule-based and frozen-model subsets are directly deployable without an external feedback channel.

Cost model. Each $\tau(s_t, a)$ is the standalone wall-clock time recorded for that arm's independently executed job in the benchmark log. Summing these costs along $\mathcal{M}(s_t)$ therefore approximates an uncached escalation workflow and does not assume artifact reuse between, for example, SP and GeoSP arms. The replay comparisons are internally consistent across policies, but the absolute cumulative cost of a cached production workflow could be lower than the standalone sums reported here.

Evaluation metrics. Per seed and benchmark the stream of committed predictions is sampled at $N_{\text{budget}} = 30$ checkpoints uniformly spaced in cumulative cost; at each we record the capped MAE and the Spearman rank correlation ρ between committed predictions and references. Per-benchmark values are seed medians; "neutral" (16 benchmarks) and "charged" (2 benchmarks) aggregates are medians across benchmarks of those seed medians. ORACLE is reported as a per-stream lower bound, not as an achievable deployment policy; no explicit cumulative-regret curve is reported.

Reproducibility. Each policy is replayed under $N_{\text{seeds}} = 10$ stream permutations (numpy.random.default_rng(seed), $seed \in \{0, \ldots, 9\}$). Uncertainty bands are non-parametric $[p_{10}, p_{90}]$ intervals over the per-seed benchmark-aggregated curves. Priors and Meta-GBMs are refit leave-one-benchmark-out per target.

Artifacts. Per-seed curves and seed-aggregated summaries live in results_sequential/sequential_results_merged.json; a flat per-benchmark view of the reported policies is results_sequential/summary_table.csv. Figures for §4: 2×2 cost–MAE/ρ learning curves (plots_sequential/sequential_bandit_2x2.png), per-benchmark MAE bars (per_benchmark_mae_bars.png), Pareto panels (pareto_neutral.png, pareto_charged.png), and top-15 dynamic-policy rankings (top15_neutral.png, top15_charged.png).

B.6. Detailed Agent Interface and Skill Packaging

Scope. As an application-layer extension, the RAPIDS oracle of §3.2 is packaged as a progressively disclosed Skill backed by a Model Context Protocol (MCP) server (Anthropic, 2024b; Model Context Protocol Contributors, 2025) so that external large language model (LLM) agents can invoke the same guarded search, validation, and escalation machinery used in the benchmark. This layer does not change the underlying energy model or any decision thresholds. Its role is to expose RAPIDS as a compact, auditable tool interface while also surfacing the multi-fidelity benchmark matrix and offline-replay artifacts that support budget-aware follow-up decisions.

Interface summary. The interface gives the agent a small number of high-level actions for live screening, higher-fidelity follow-up, run inspection, and replay-based planning, with detailed instructions disclosed only when needed. In practice this keeps the prompt footprint small while still allowing the agent to access the nine-fidelity benchmark summaries and the replay-policy artifacts derived from the same $5{,}565 \times 9$ log.

Client runtime and permissions. In the configuration used for this paper, the Skill is installed in Claude Code and exercised with a Claude Sonnet 4.5 client agent (Anthropic, 2025b). That agent is granted direct permission to call the RAPIDS MCP server, so live binding screening, DFT or CREST follow-up, and replay-policy queries are executed through the tool interface rather than as natural-language approximation. This statement is about the agent runtime and permission model only; it does not change the scientific pipeline defined in §3.2–§3.4.

Prompt harnesses for the PFAS downstream demonstration. For the PFAS-oriented downstream study of §4.7, we instantiated the prompt primitive as an explicit *harness*: a preset task specification, analyte set, allowed tools, and reporting schema. The PFAS demonstration uses three agent configurations of six rollouts each.

Single-agent harness. A single Claude Sonnet 4.5 agent receives the coordinator instructions below and produces all probe evaluations within one context window without delegating to sub-agents. This is the lowest-budget configuration (typically 5–6 probes per rollout).

Meta-agent / sub-agent harness. The top-level Claude Code agent acts as a coordinator and, when a probe deserves dedicated analysis, dispatches a worker sub-agent for that probe–analyte pair. The coordinator prompt is: "You are comparing four neutral analytes, PFOS, PFOA, DDS, and TCAA, while exploring multiple candidate probes with the RAPIDS Skill. The goal is to determine whether PFOS and PFOA are preferred over DDS and TCAA under the RAPIDS binding score. You may invoke the RAPIDS Skill, inspect run artifacts, request escalation when confidence is low or disagreement is large, and, if available, launch sub-agents to analyze individual analyte–probe pairs. Base the final decision only on MCP-returned evidence. Report a ranked table, the run identifiers used, the fidelity reached for each analyte, and a short explanation of why the target-versus-interferent ordering is or is not supported." The worker sub-agent prompt is: "You are assigned one analyte–probe pair inside the PFAS harness. Start from RAPIDS screening, inspect confidence and guard outcomes, and escalate only if the RAPIDS result is flagged, low-confidence, or materially ambiguous for the final ranking task. Return the committed score, confidence label, guard summary, run identifier, and a short note on whether this pair appears favorable or unfavorable relative to the harness objective. Do not use hidden labels, backend scripts, or external chemistry heuristics beyond the MCP-exposed outputs." This configuration enables broader probe coverage (typically ~ 16 probes per rollout).

Meta-agent / sub-executor-agent / sub-critic-agent harness. (Barrak, 2025; Shinn et al., 2023) The coordinator dispatches two distinct sub-agent roles. The *sub-executor agent* proposes and evaluates probes following the worker harness above. The *sub-critic agent* observes the executor's running probe set and per-target binding energies, computes intermediate Δ_{select}, and returns structured suggestions; its prompt is: "You are a critic agent in a PFAS-probe discovery loop. Given the executor's current probe set and PFOS/PFOA/DDS/TCAA binding energies, identify (1) the top current Δ_{select} winner, (2) chemically adjacent neighbors to probe next (local analog expansion), (3) motif families currently underexplored (orthogonal expansion), and (4) any counterexample probes worth checking. Return ranked suggestions and short rationale only; do not run RAPIDS or call any tool directly." The coordinator routes the critic's suggestions back to the executor for follow-up. This is the highest-budget configuration (typically ~ 25–$30+$ probes per rollout, with $30+$ MCP tool calls). We report these harnesses in abridged form because the scientific point is the task framing and tool-budget separation across the three configurations; we omit the exact natural-language phrasing.

Relation to the benchmark. The MCP layer exposes two scientifically distinct capabilities. First, it exposes the *oracle path*: direct RAPIDS screening with optional DFT or CREST fallback, which is the pathway benchmarked in the results section. Second, it exposes the *policy path*: access to the precomputed replay log and its learned allocation policies, which allows an external agent to ask for a budget-aware fidelity recommendation before launching any expensive follow-up. In practice, the interface wraps both the RAPIDS predictor and the empirical decision policy learned from the $5{,}565$-system benchmark. In downstream use this packaging supports agent-mediated probe-screening loops, including OOD tasks such as PFAS sensing. The empirical agent evidence in this paper remains limited to the underlying oracle and replay policies together with a bounded PFAS-oriented harness.

Selectivity metric. Probe selectivity is scored by

$$\Delta_{\text{select}} = \tfrac{1}{2}\big(\Delta E_{\text{PFOS}} + \Delta E_{\text{PFOA}}\big) - \tfrac{1}{2}\big(\Delta E_{\text{DDS}} + \Delta E_{\text{TCAA}}\big),$$

where ΔE_X is the probe–X relaxed binding energy. More negative Δ_{select} indicates stronger PFAS (PFOS+PFOA) preference over the interferents (DDS+TCAA). All four-analyte binding energies in the rollout figures of Appendix C.8 are evaluated at the same fidelity rung, ωB97M-V/def2-TZVPD GeoSP, so Δ_{select} comparisons across probes and across rollouts are at uniform fidelity.

Rollout coverage across configurations. For the PFAS case study summarized in §4.7, the single-agent configuration covers roughly 5–6 probes per rollout; meta-agent / sub-agent rollouts cover ~ 16 probes per rollout; meta-agent / sub-executor / sub-critic rollouts cover $\sim$ 25–30+ probes per rollout, with the longest individual rollout exceeding 30 MCP tool calls. The combined probe set across the 18 rollouts spans cationic small-molecule scaffolds, host-cavity macrocycles, fluorophilic anchors, lactam/amide donors, and alkyl/perfluoroalkyl ammonium tail-engineering motifs.

Top-25 recovery and motif analysis. Aggregating the 18 rollouts against the uploaded master table, the single-agent, meta-agent / sub-agent, and meta-agent / sub-executor / sub-critic configurations recover 8, 17, and 25 of the global top-25 Δ_{select} probes, respectively; the same monotonic improvement holds for the top-10 set and for strict PFAS-over-interferent winners. The leading edge of the search space is dominated by compact cationic bicyclic and quaternary-ammonium scaffolds: N-quinuclidinium ($\Delta_{\text{select}} = -18.19$ kcal/mol), N-methyl-DABCO (-14.30), 1-methyl-DABCO (-14.23), N-propyl-quinuclidinium (-14.07), granatanine-quaternized (-13.04), and butyrylcholine (-12.95), all strict PFAS-over-interferent winners. Among more fully evaluated anchors, γ-cyclodextrin remains the strongest neutral benchmark (-12.16, $n_{\text{real}} = 4$), followed by 1,1,3,3-tetramethyl-guanidinium (-11.39) and protonated 9-ABN (-11.35). The second tier includes triphenyl-sulfonium, tetramethylammonium, guanidinium, 1-octylpyridinium, 1-methyl-3-perfluorohexyl-imidazolium, protonated TBD, 1-methyl-3-perfluorooctyl-imidazolium, decyltrimethylammonium, perfluorooctyl-trimethylammonium, cucurbit-6-uril, dodecyltrimethylammonium, choline, N,N,N-trimethyl-perfluorohexyl-ammonium, propyl-triphenyl-phosphonium, N-methyl-tropanium, and perfluoro-tripodalamide-3-mesitylene. Three motif-level patterns emerge: (i) rigid compact cationic bicyclic / quaternary-ammonium scaffolds dominate the top of the table; (ii) host-type cavities (cyclodextrin, cucurbit-N-uril) remain the most credible neutral direction; (iii) once a favorable cationic head group is found, alkyl/perfluoroalkyl tail engineering further sharpens PFAS bias, populating the second tier.

Why the critic-augmented configuration wins, with one caveat. The meta-agent / sub-executor / sub-critic configuration wins not by evaluating more compounds in isolation, but by adaptively expanding emerging winners (local analog sweeps) and revisiting orthogonal motif families flagged by the critic (mode-collapse mitigation). The caveat is that some critic-augmented rollouts deliberately include exploratory and counterexample probes, so not every individual ultra-high rollout has a higher mean Δ_{select} than every meta-agent / sub-agent rollout; the cleanest summary is therefore that this configuration is the best *discovery* policy, in that it recovers the broadest and strongest PFAS-selective region of chemical space, even though some individual long rollouts contain exploratory or confirmatory compounds with weaker mean selectivity.

C. Supplementary results figures

C.1. Full 18-benchmark fidelity overview

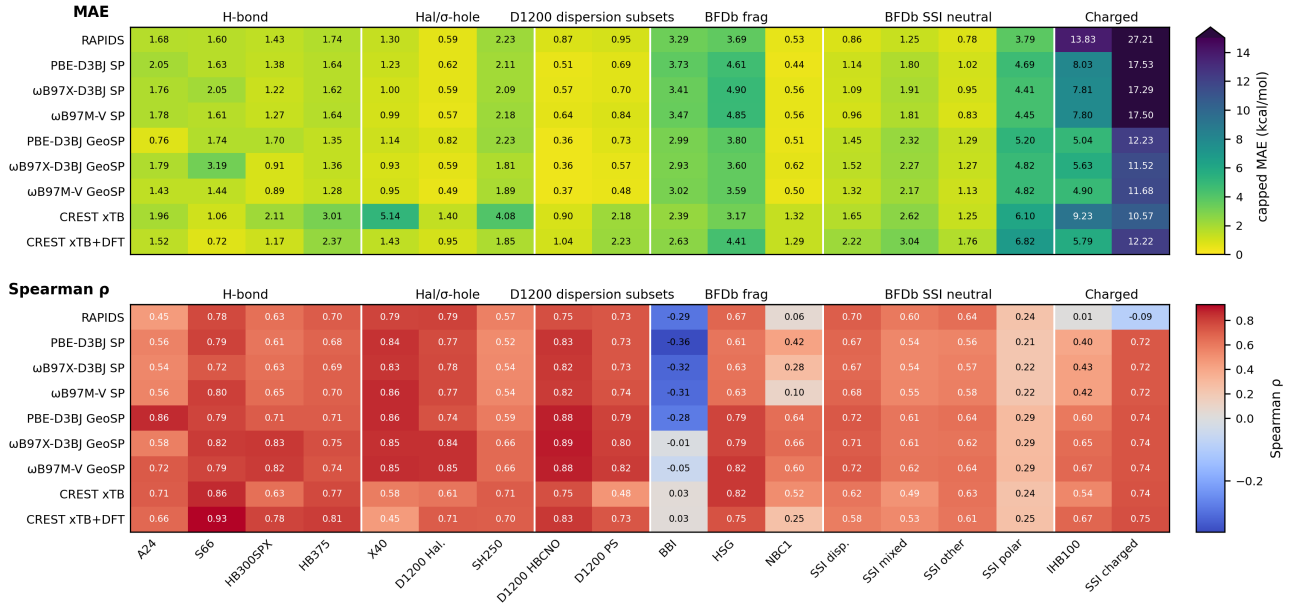


Figure S1. Full 18-benchmark benchmark-relative summary of the nine fidelity methods. Top panel: capped MAE (kcal/mol). Bottom panel: Spearman ρ. The main text concentrates on the neutral story and defers these full per-benchmark heatmaps, including the two charged benchmarks, to the appendix.

C.2. Charged-system diagnostics

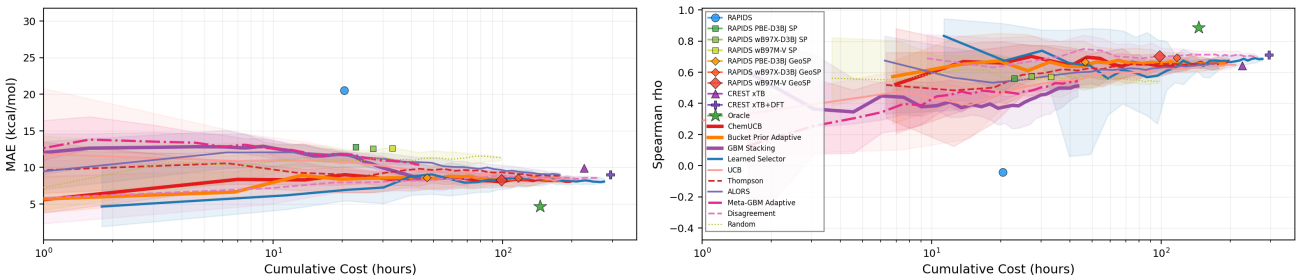


Figure S2. Charged-only offline-replay trajectories deferred from the main text. Left panel: capped MAE versus cumulative cost, in hours on a log scale. Right panel: Spearman ρ versus cumulative cost. This is the charged counterpart of the main-text neutral replay figure and shows the same qualitative pattern discussed in §4.5: RAPIDS direct output is not competitive on charged motifs, while DFT geometry correction and CREST-based sampling recover substantially better MAE and ρ.

C.3. Error-distribution diagnostics

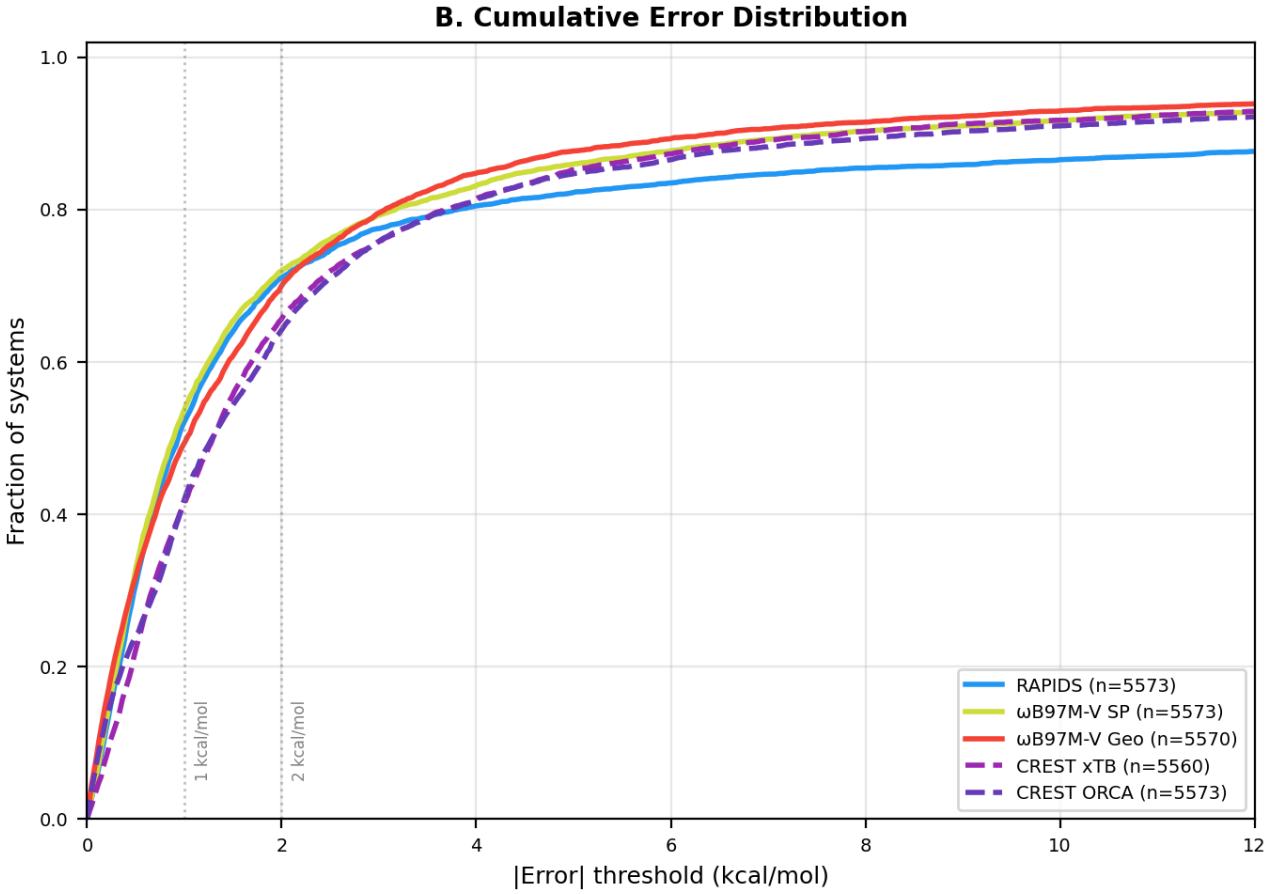


Figure S3. Distribution-level diagnostics deferred from the main text. Cumulative distribution of absolute errors for RAPIDS, ωB97M-V/def2-TZVP SP, ωB97M-V GeoSP, CREST xTB, and CREST xTB+DFT, aggregated over all available systems. This panel shows that RAPIDS remains competitive in the low-error regime but develops a heavier tail than the upgraded DFT and CREST baselines.

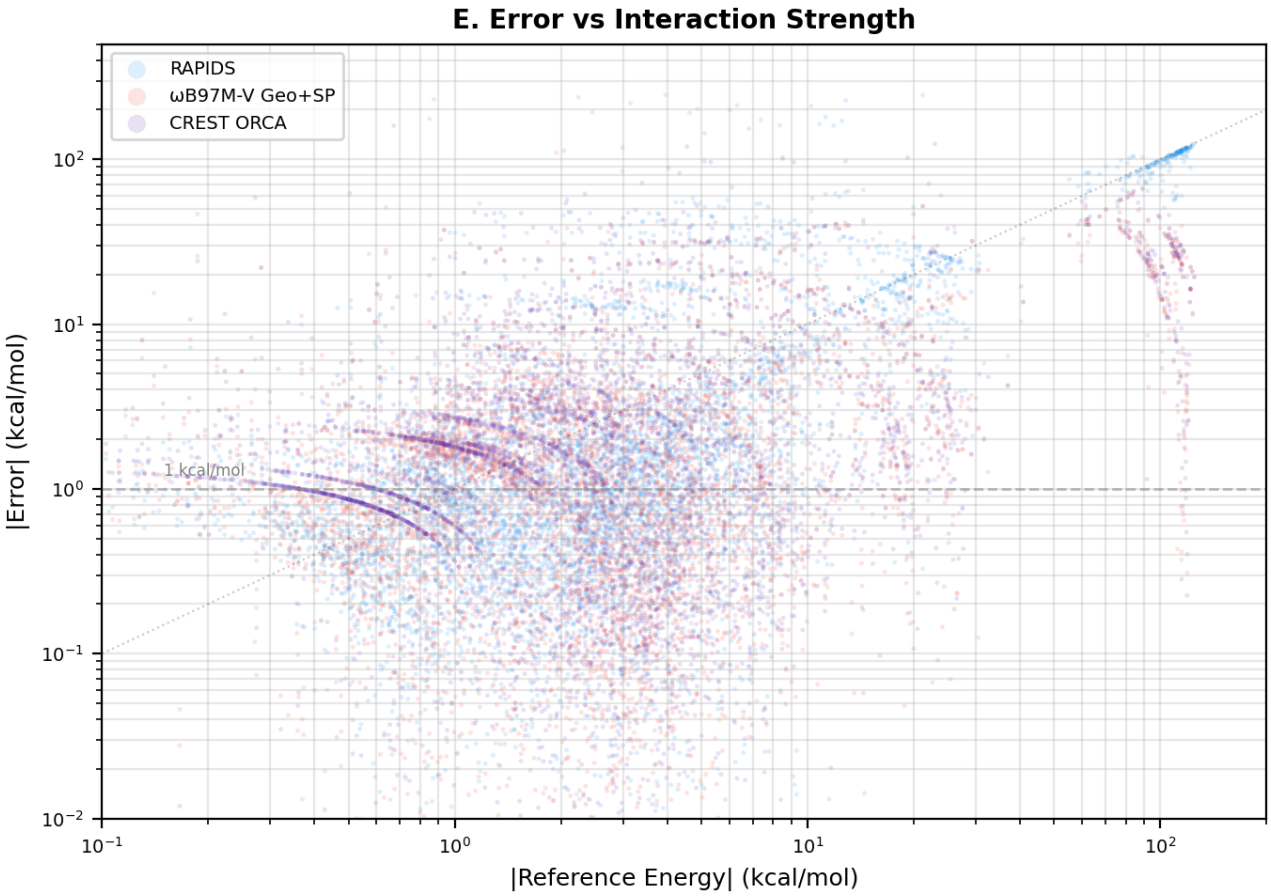


Figure S4. Absolute error versus absolute reference interaction energy for RAPIDS, ωB97M-V GeoSP, and CREST xTB+DFT. This panel complements Fig. S3 by showing how the RAPIDS error tail expands as interaction strength and charged-character increase, while the upgraded DFT and CREST baselines remain better controlled.

C.4. Static aggregate summaries

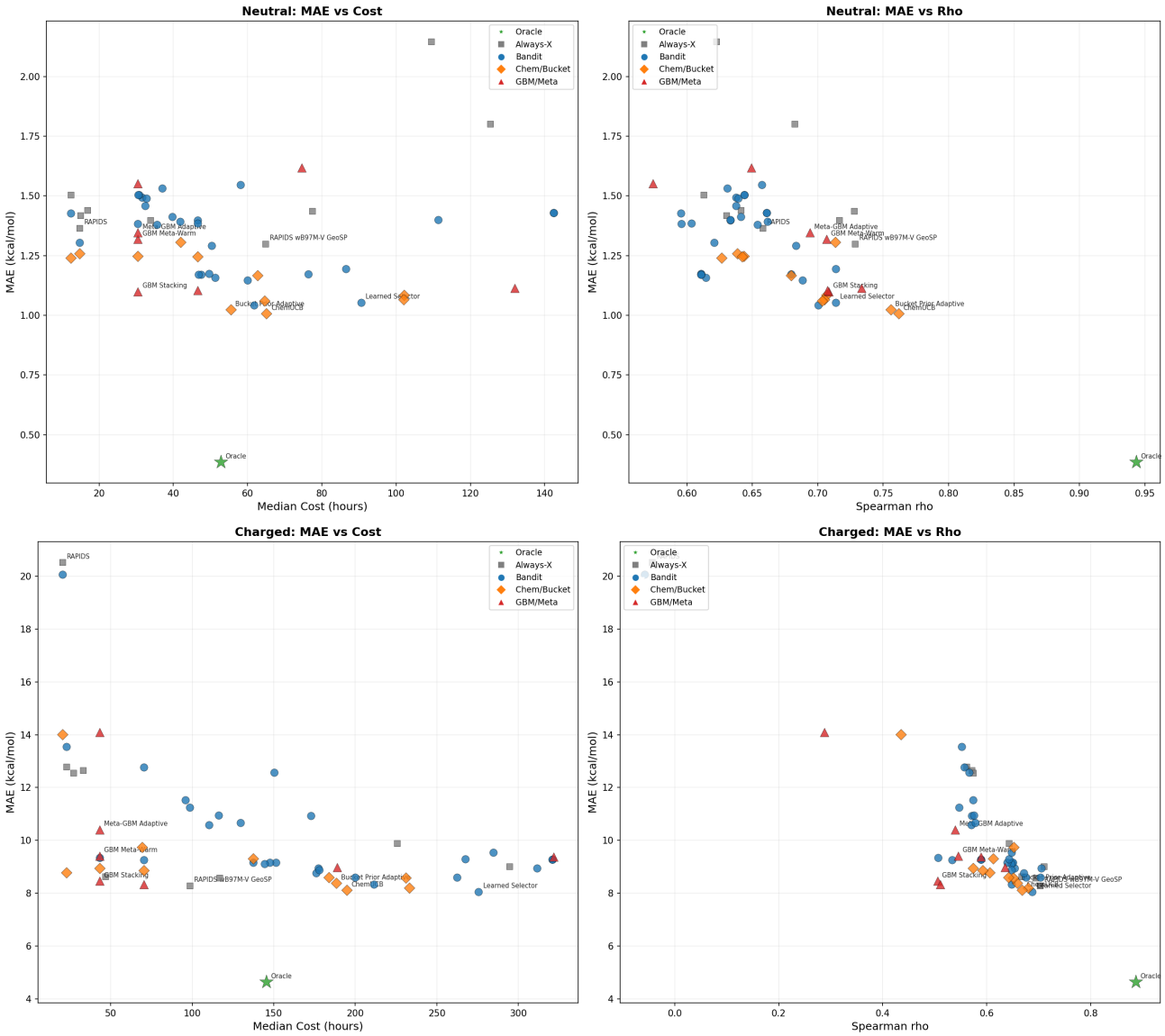


Figure S5. Aggregate cost–accuracy summaries deferred from the main text. Top: neutral aggregate. Bottom: charged aggregate. In each panel, the left subplot shows capped MAE versus aggregate cost and the right subplot shows Spearman ρ versus MAE. Gray squares are the fixed single-fidelity baselines; the remaining colored markers are the 40 reported replay policies, colored by policy family; ORACLE gives the per-system lower bound. These summaries complement the main-text neutral static-fidelity overview by retaining the aggregate view for both charge classes.

C.5. Policy-family rankings

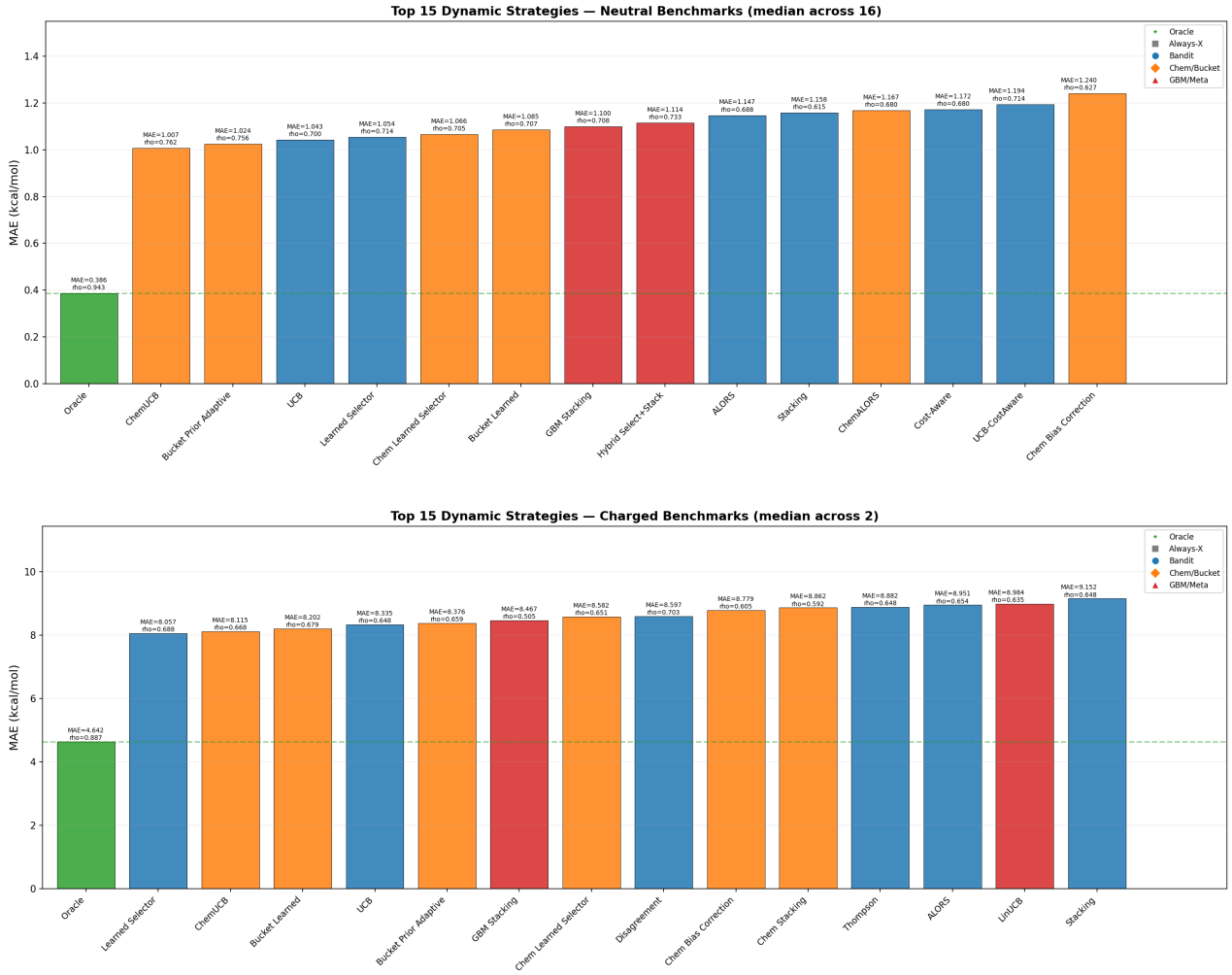


Figure S6. Top-15 replay policies by final capped MAE, neutral aggregate (top) and charged aggregate (bottom). Static ALWAYS-X policies and ORACLE are shown for reference. The policy families that rank near the top differ between the two aggregates.

C.6. Representative benchmark panels

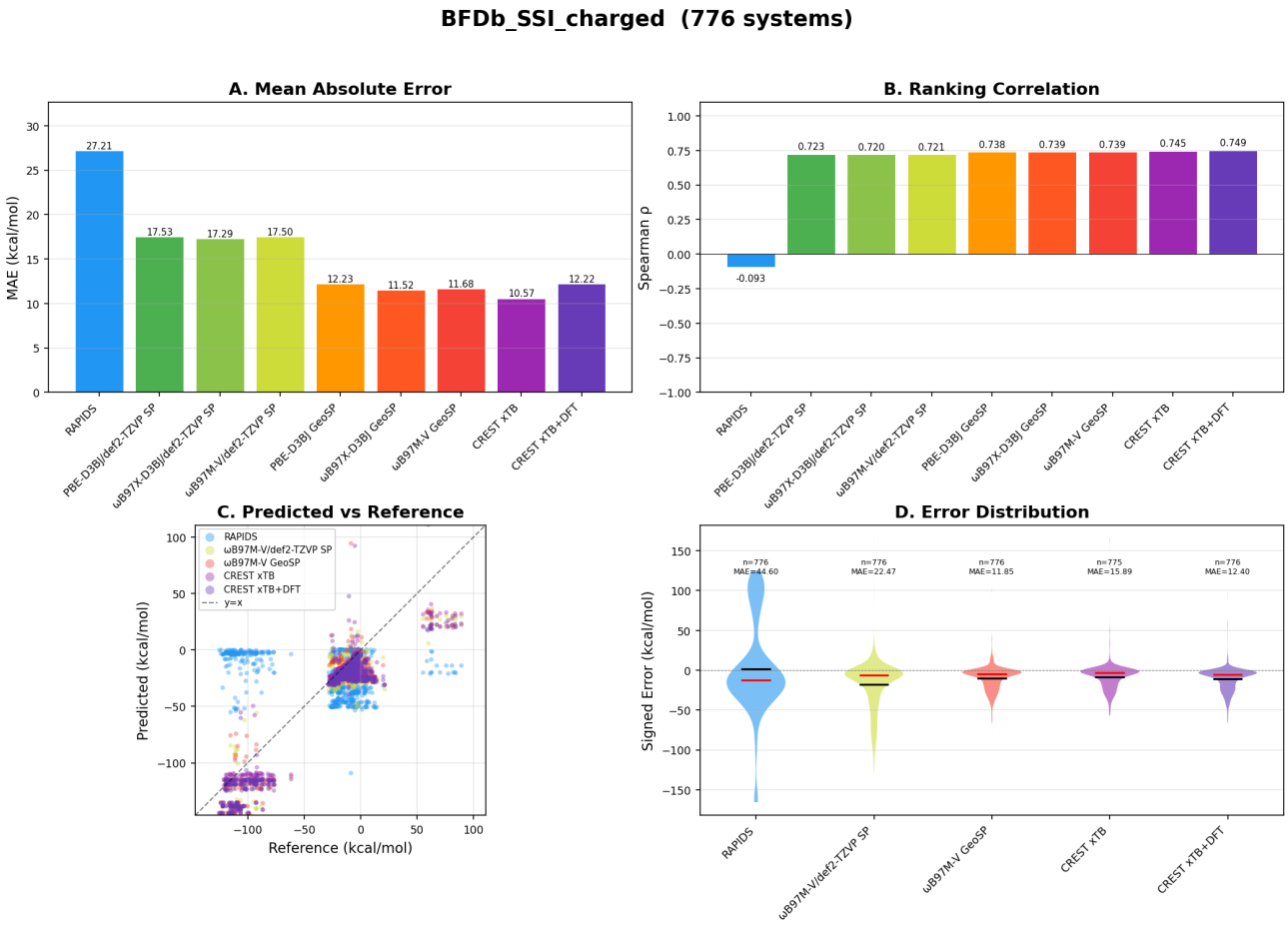


Figure S7. Representative charged benchmark: BFDb SSI charged. The four subpanels report method-wise MAE, method-wise Spearman ρ, predicted-versus-reference scatter for representative methods, and signed-error distributions. This case makes the charged failure mode concrete: RAPIDS direct output is poor, while DFT geometry correction and CREST-based sampling recover much stronger ranking behavior and lower MAE.

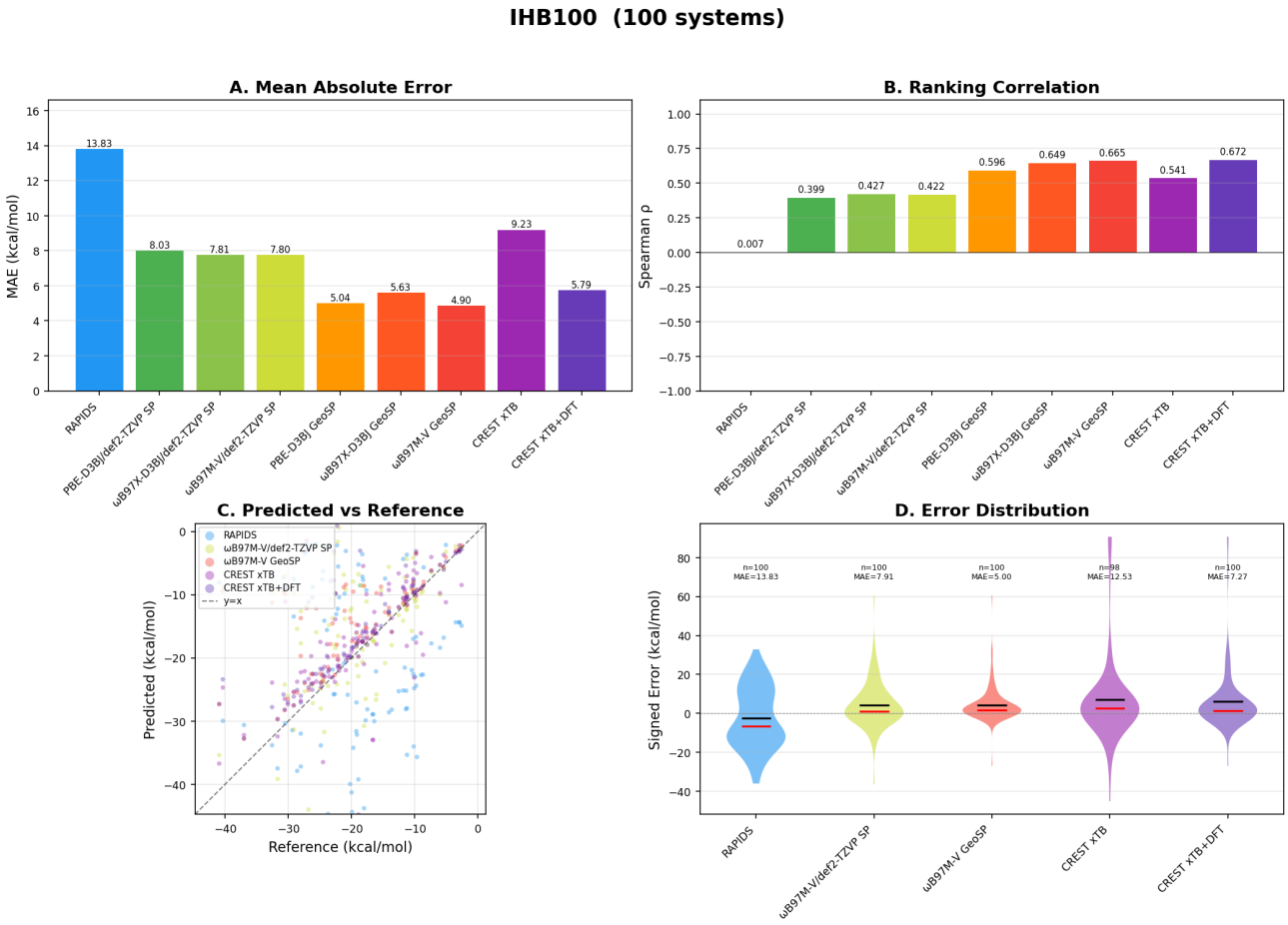


Figure S8. Representative charged benchmark: IHB100. The four subpanels report method-wise MAE, method-wise Spearman ρ, predicted-versus-reference scatter for representative methods, and signed-error distributions. This benchmark shows the same qualitative pattern as BFDb SSI charged, but with a clearer recovery once the RAPIDS geometry is upgraded explicitly at the DFT level.

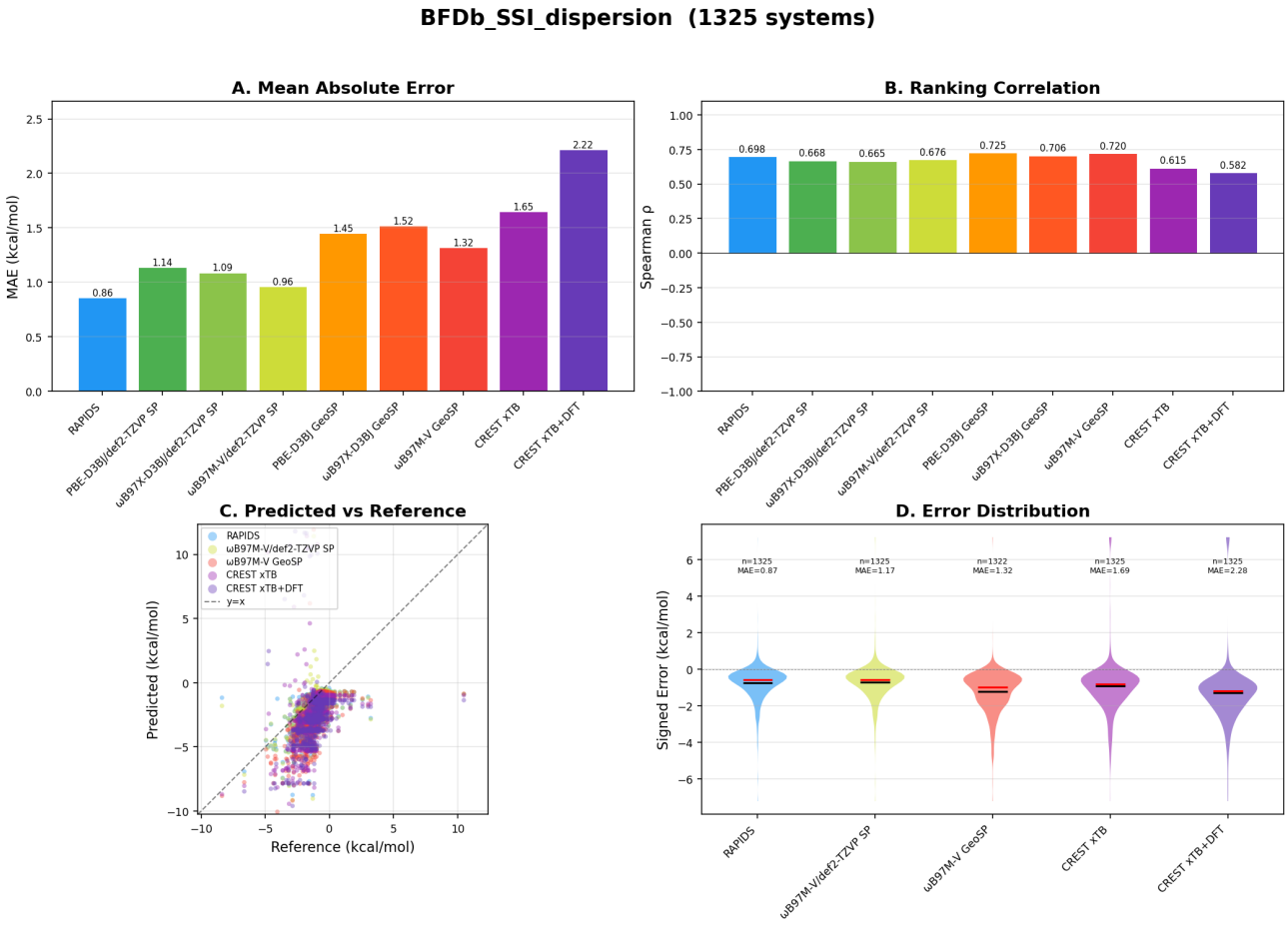


Figure S9. Representative neutral benchmark: BFDb SSI dispersion. The four subpanels report method-wise MAE, method-wise Spearman ρ, predicted-versus-reference scatter for representative methods, and signed-error distributions. This is the benchmark-relative neutral setting in which RAPIDS is strongest in the main text.

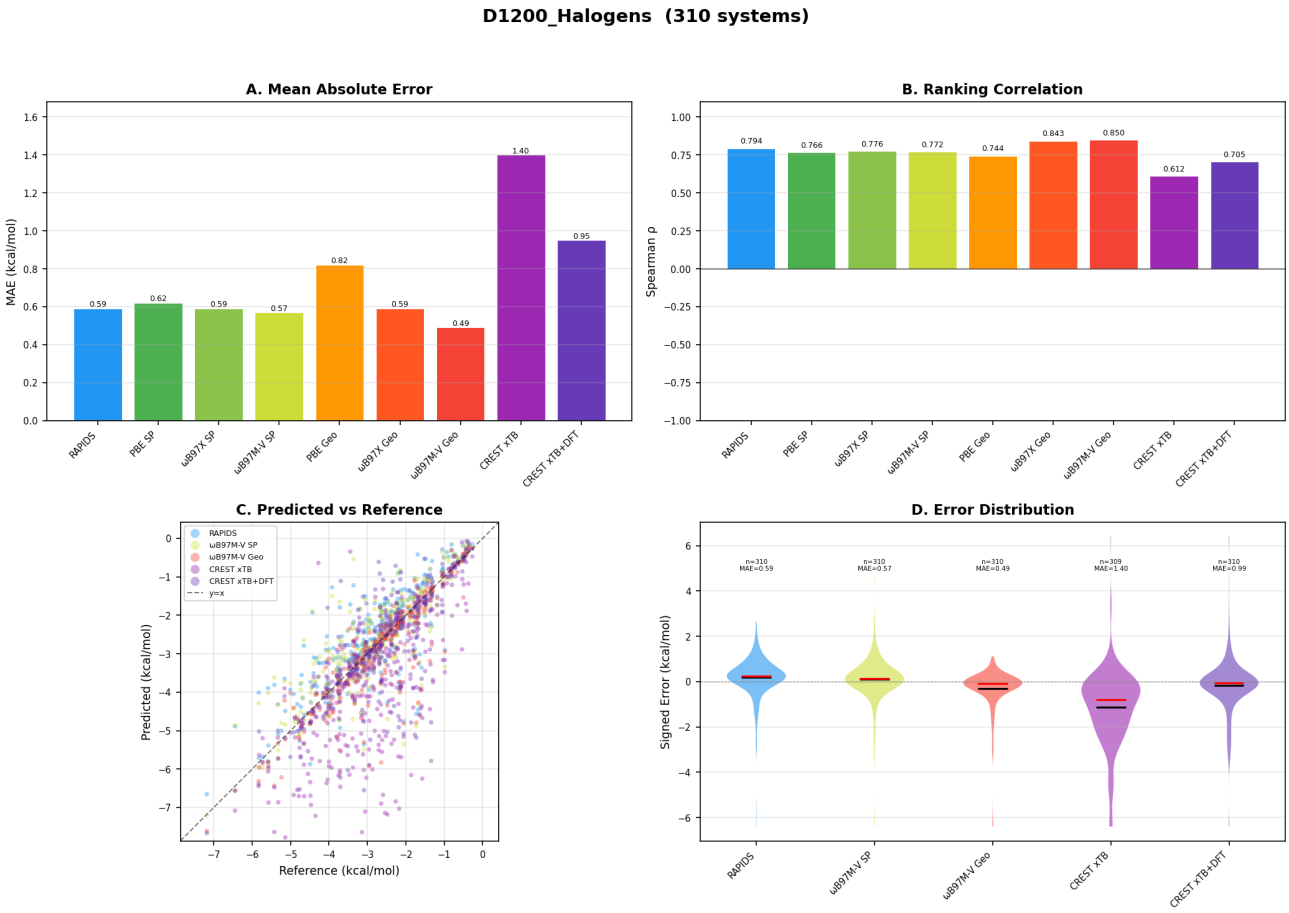


Figure S10. Representative neutral benchmark: D1200 halogens. The four subpanels report method-wise MAE, method-wise Spearman ρ, predicted-versus-reference scatter for representative methods, and signed-error distributions. This panel shows a complementary neutral case in which explicit DFT and CREST variants regain the lead over RAPIDS.

C.7. Cross-backbone diagnostic

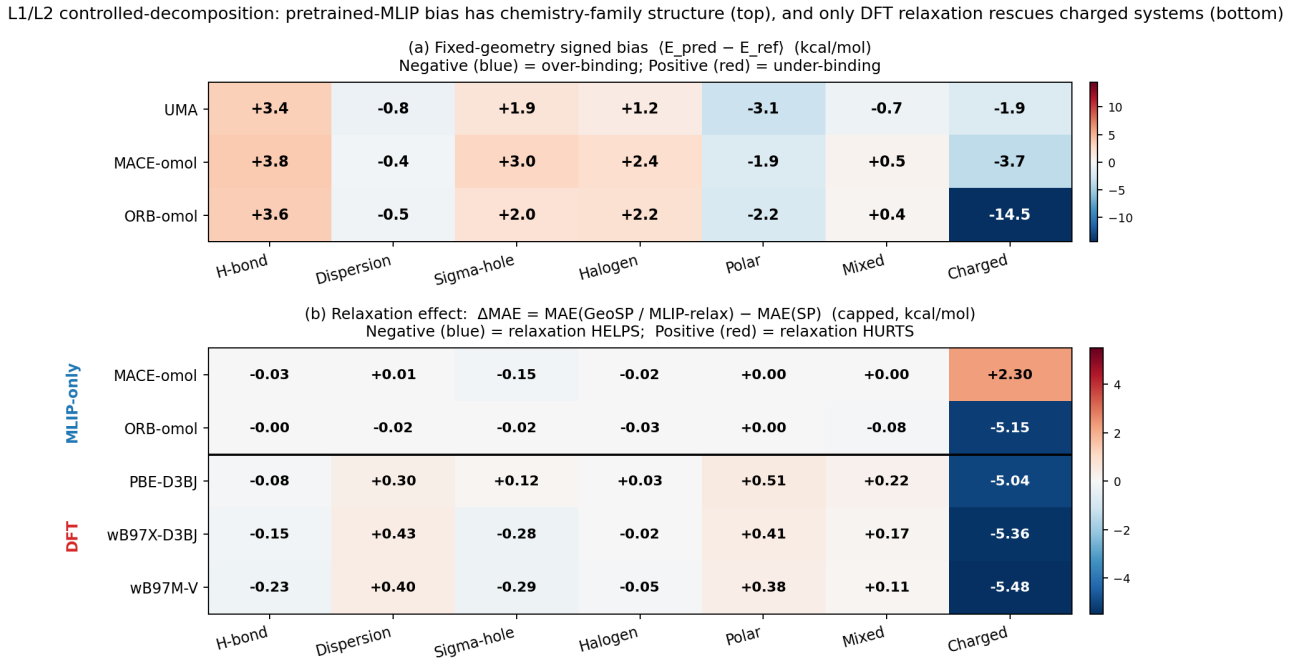


Figure S11. Cross-backbone diagnostic on RAPIDS-committed geometries. (a) Signed bias $\langle E_{\text{pred}} - E_{\text{ref}} \rangle$ by chemistry family for UMA, MACE-omol, and ORB-omol fixed-geometry single points (red = under-binding, blue = over-binding). (b) Relaxation effect $\Delta\text{MAE} = \text{MAE(GeoSP/relax)} - \text{MAE(SP)}$ per backbone and family (red = relaxation hurts, blue = helps). Interpretation in §4.6.

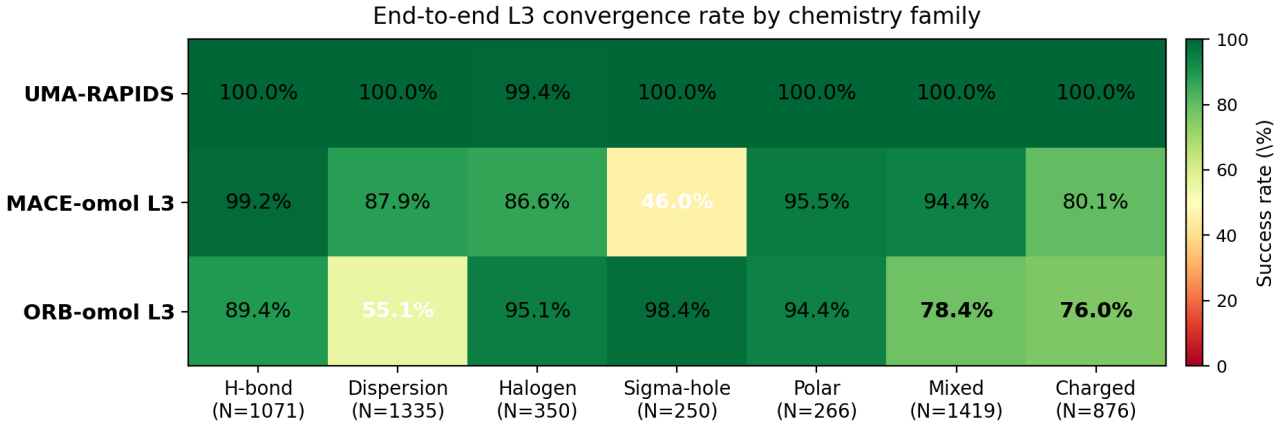


Figure S12. L3 end-to-end convergence rate by chemistry family. UMA-RAPIDS is the universal-coverage baseline; MACE-omol L3 and ORB-omol L3 have complementary failure pockets (details below).

End-to-end L3 results. On the fair subset where both alternative L3 runs converge ($N = 4{,}041$), ORB-omol relaxation reaches capped MAE 2.75 kcal/mol against ωB97M-V GeoSP's 2.77 (statistical parity); ORB-omol L3 reaches 3.01 and MACE-omol L3 3.33, all well below UMA-RAPIDS at 5.78 on the same subset. Coverage is asymmetric (Fig. S12): MACE-omol converges on 0% of X40 and 46% of SH250 (heavy-halogen and σ-hole motifs), whereas ORB-omol converges on only 55% of BFDb_SSI_dispersion. Across the full 5,565-system stream, MACE-omol succeeds where ORB-omol fails on 907 systems and ORB-omol succeeds where MACE-omol fails on 260, with 357 systems on which both alternative backbones fail. ORB-omol's apparent parity with DFT-GeoSP is therefore coverage-conditional rather than a global replacement: UMA-RAPIDS remains the main pipeline because it covers the full stream, while MACE-omol and ORB-omol are usable as alternative-backbone arms with explicit chemistry-family exclusions. The 907/260 convergence-outcome

asymmetry also provides a deployment-side complementary signal to the prediction-level disagreement signal of §4.6: escalation can be triggered by which backbone converges, in addition to the magnitude of prediction disagreement.

Cross-backbone BiasCorr universality. The structured-bias view of finding (i) generalizes across backbones. Applying the per-benchmark BiasCorr protocol of §4.4 cross-backbone reduces neutral MAE by 8–10% at L1/L2 and substantially more at L3: MACE-omol L3 drops by 31% and ORB-omol L3 by 26%, roughly four times the relative improvement on UMA-RAPIDS under the same evaluation. A chemistry-family-grouped variant (ChemBiasCorr) further closes the gap, bringing UMA-RAPIDS, MACE-omol L3, and ORB-omol L3 neutral MAE to roughly 1.28–1.30 kcal/mol, approaching ωB97M-V GeoSP. The chemistry-family-structured bias underlying finding (i) is thus not UMA-specific and is disproportionately learnable on end-to-end alternative-backbone arms.

C.8. Agent rollout galleries

The 18 rollouts of §4.7 are partitioned into three agent configurations of 6 rollouts each: single-agent (Figs. S13–S18), meta-agent / sub-agent (Figs. S19–S24), and meta-agent / sub-executor-agent / sub-critic-agent (Figs. S25–S30). Each cell within a rollout sheet shows the geometry-reoptimized probe–target complex for one of {PFOS, PFOA, DDS, TCAA} with its annotated binding energy in kcal/mol; the sequence index "#n" indicates rollout order.

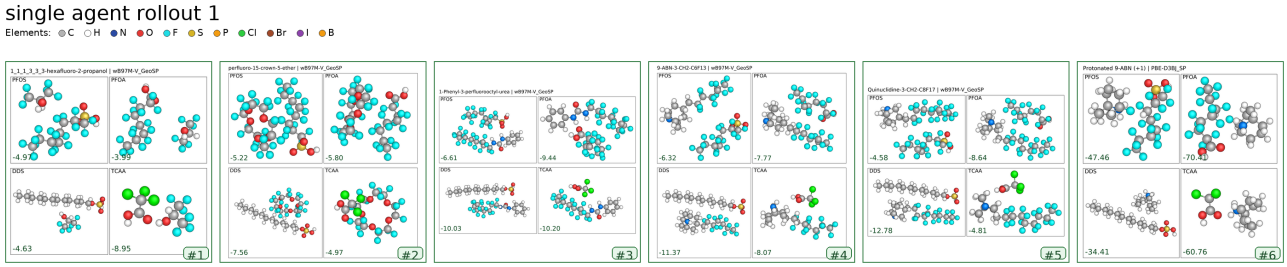


Figure S13. Single-agent rollout 1.

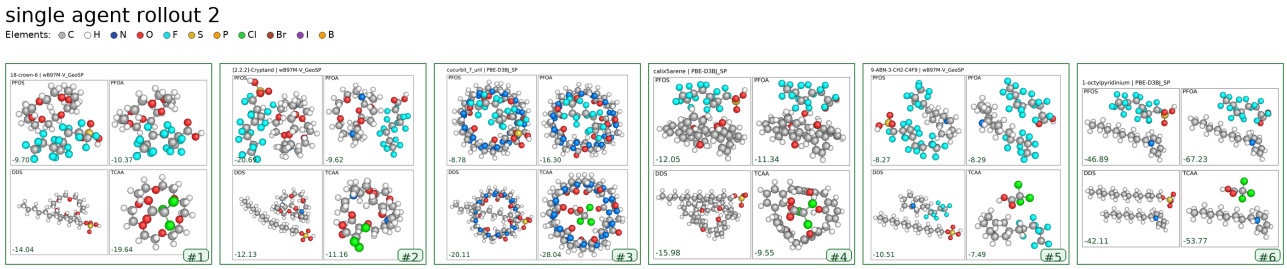


Figure S14. Single-agent rollout 2.

Single-agent rollouts (1–6).

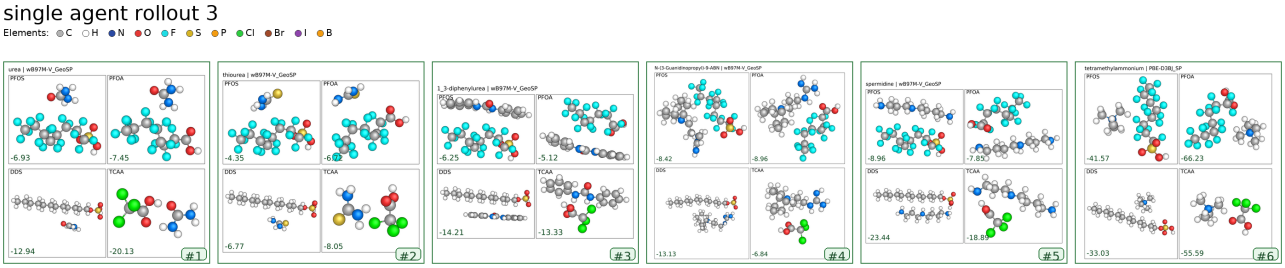


Figure S15. Single-agent rollout 3.

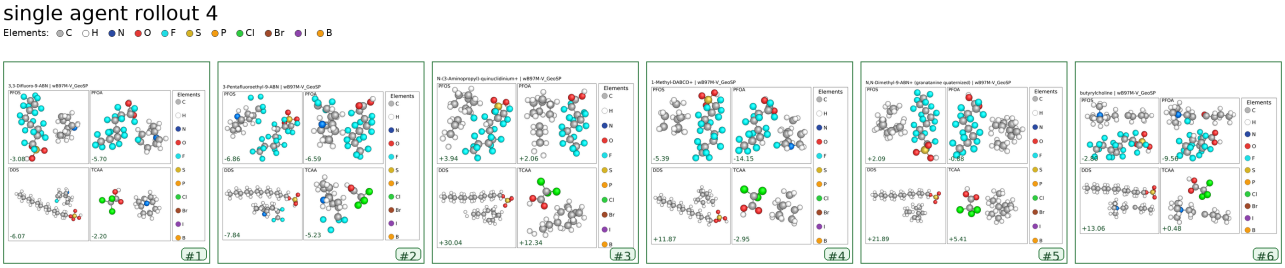


Figure S16. Single-agent rollout 4.

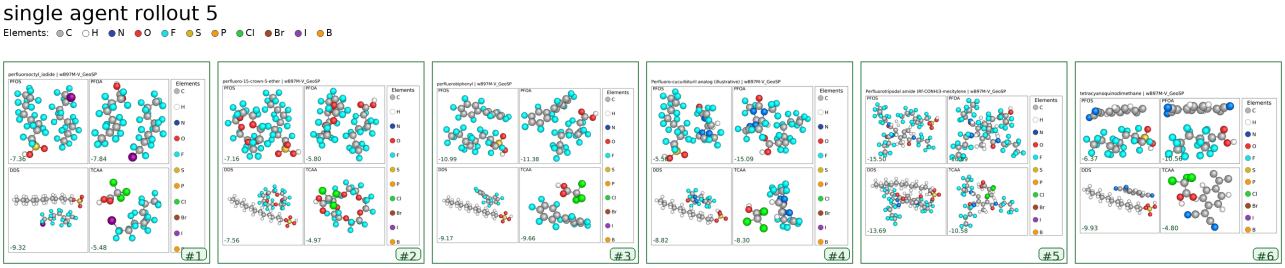


Figure S17. Single-agent rollout 5.

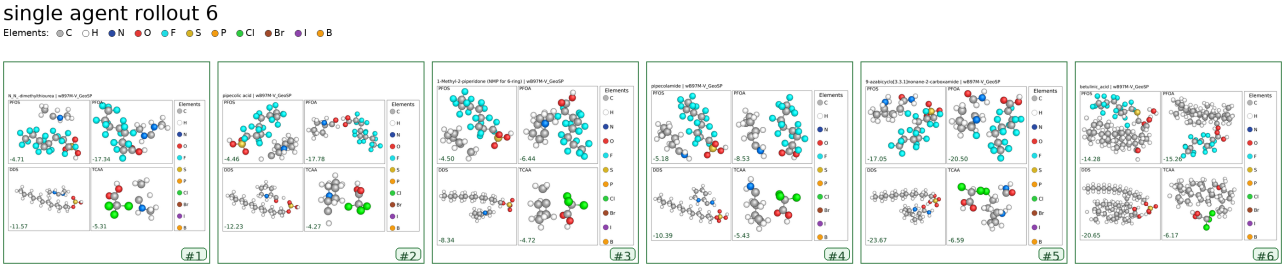


Figure S18. Single-agent rollout 6.

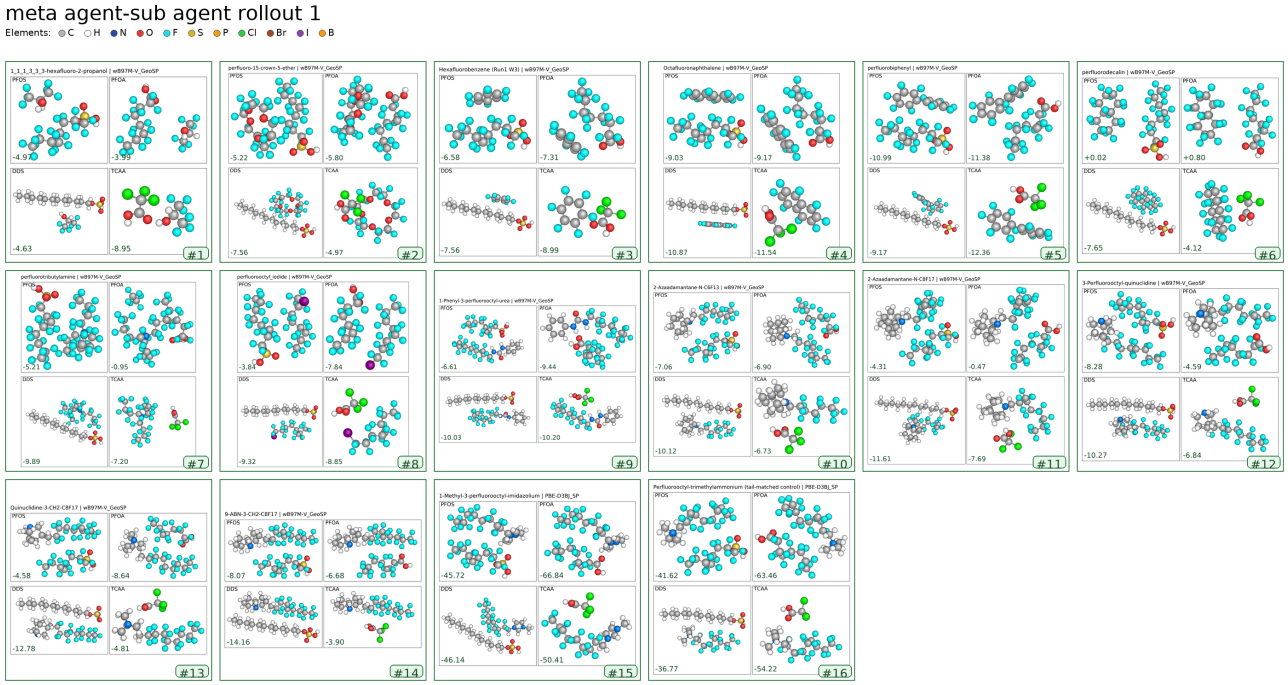


Figure S19. Meta-agent / sub-agent rollout 1.

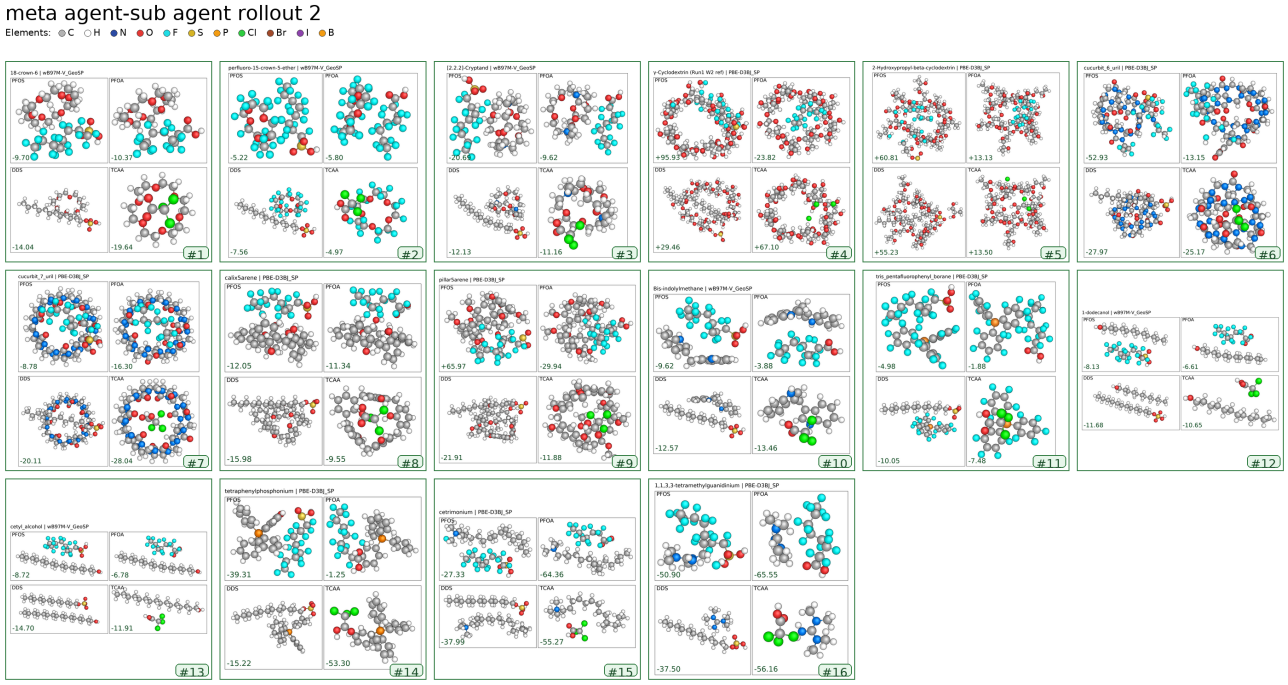


Figure S20. Meta-agent / sub-agent rollout 2.

Meta-agent / sub-agent rollouts (1–6).

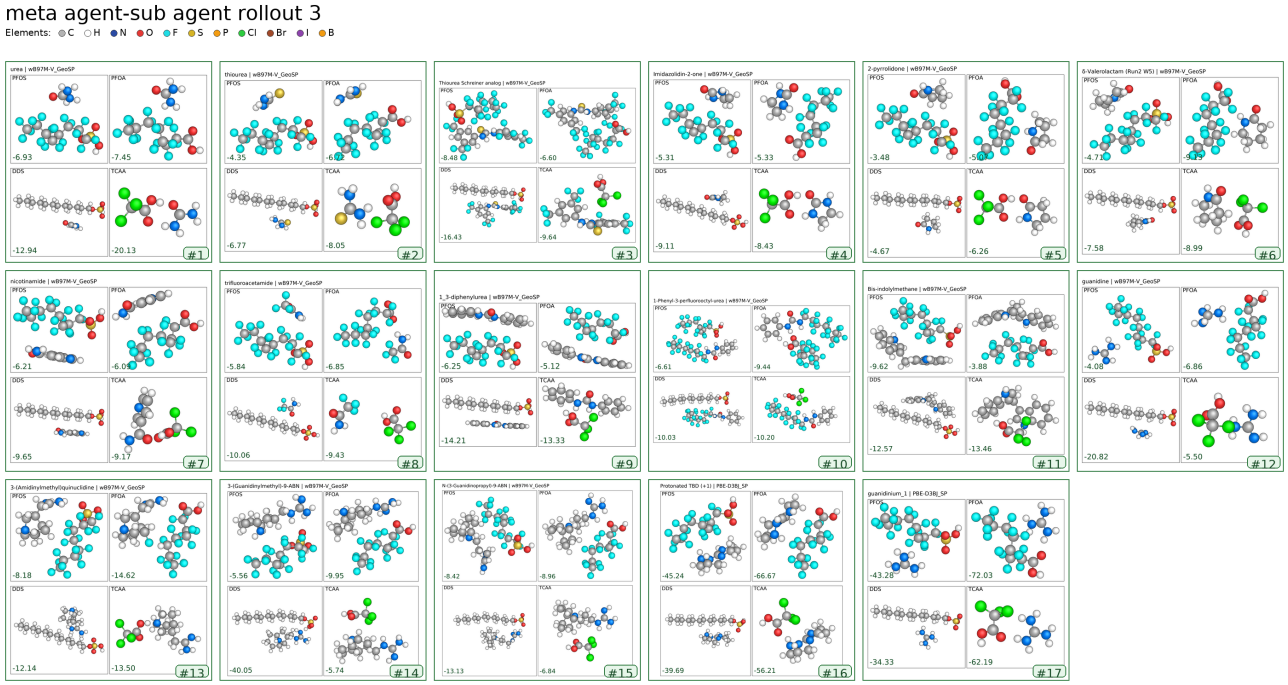


Figure S21. Meta-agent / sub-agent rollout 3.

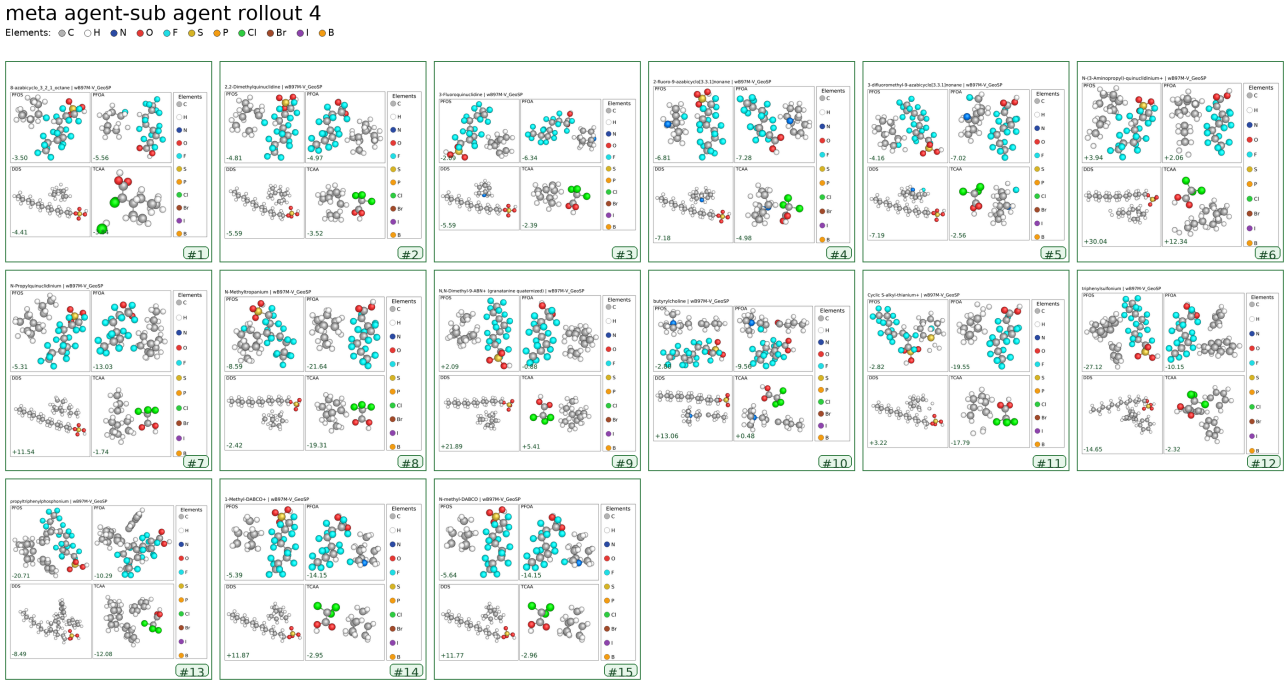


Figure S22. Meta-agent / sub-agent rollout 4.

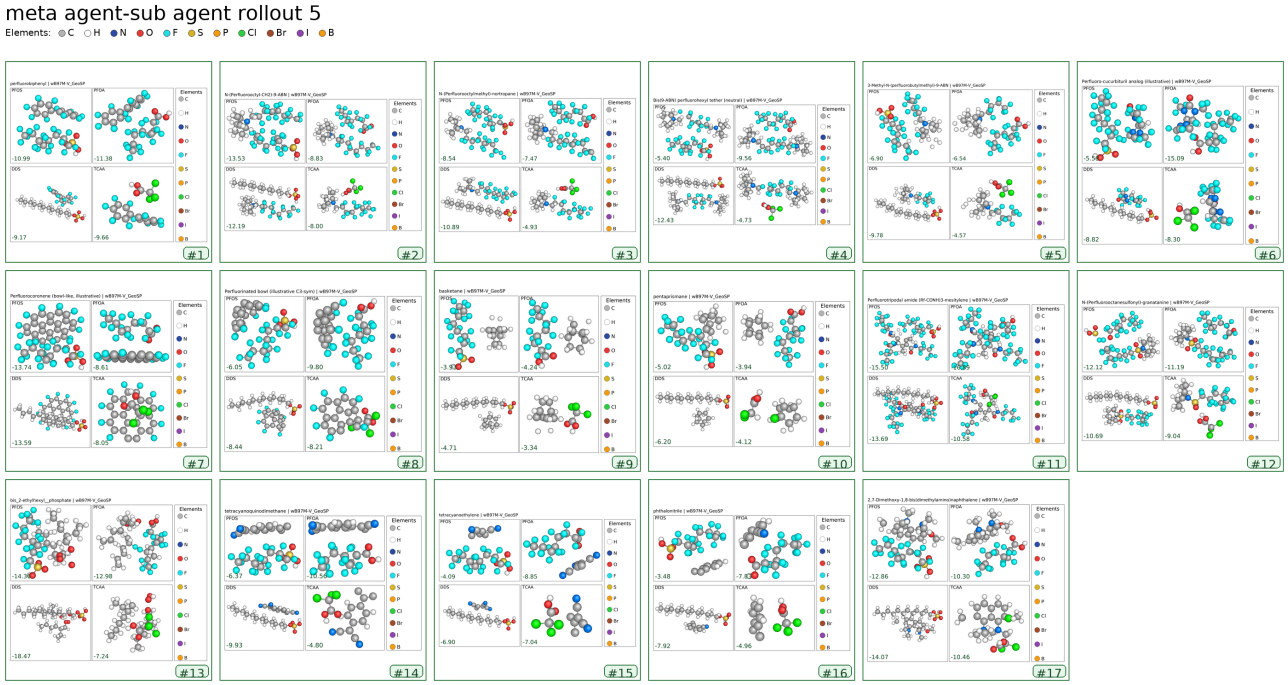


Figure S23. Meta-agent / sub-agent rollout 5.

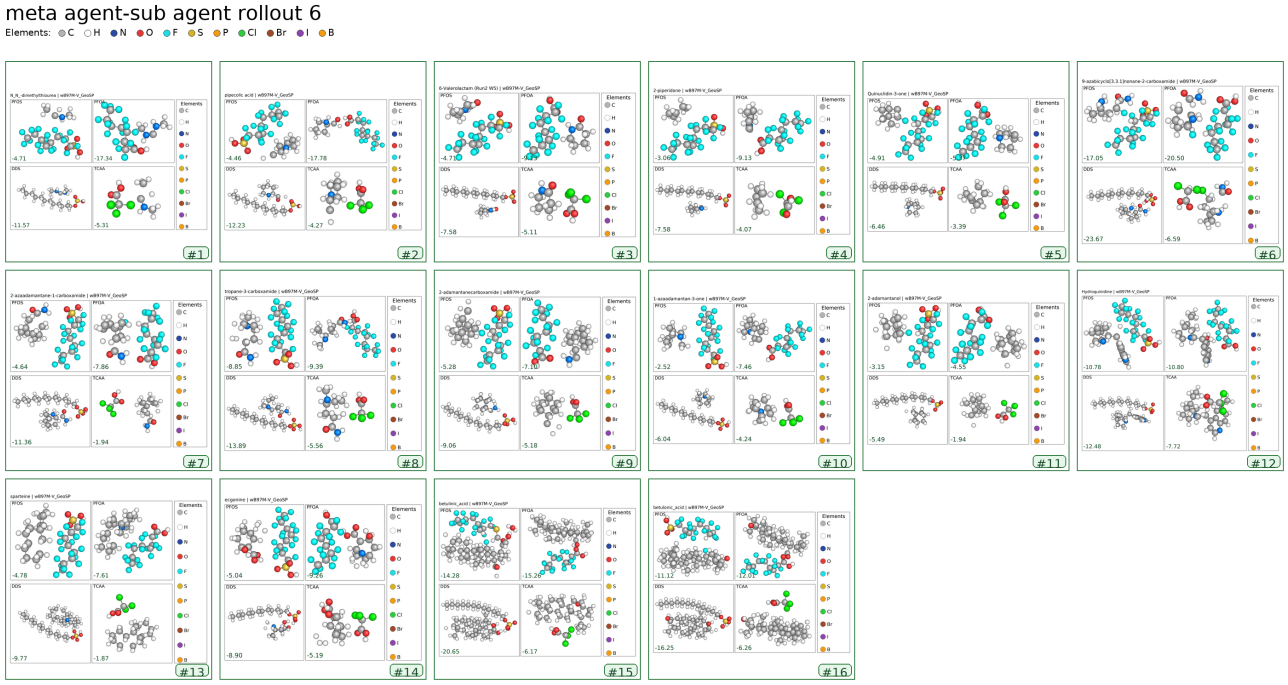


Figure S24. Meta-agent / sub-agent rollout 6.

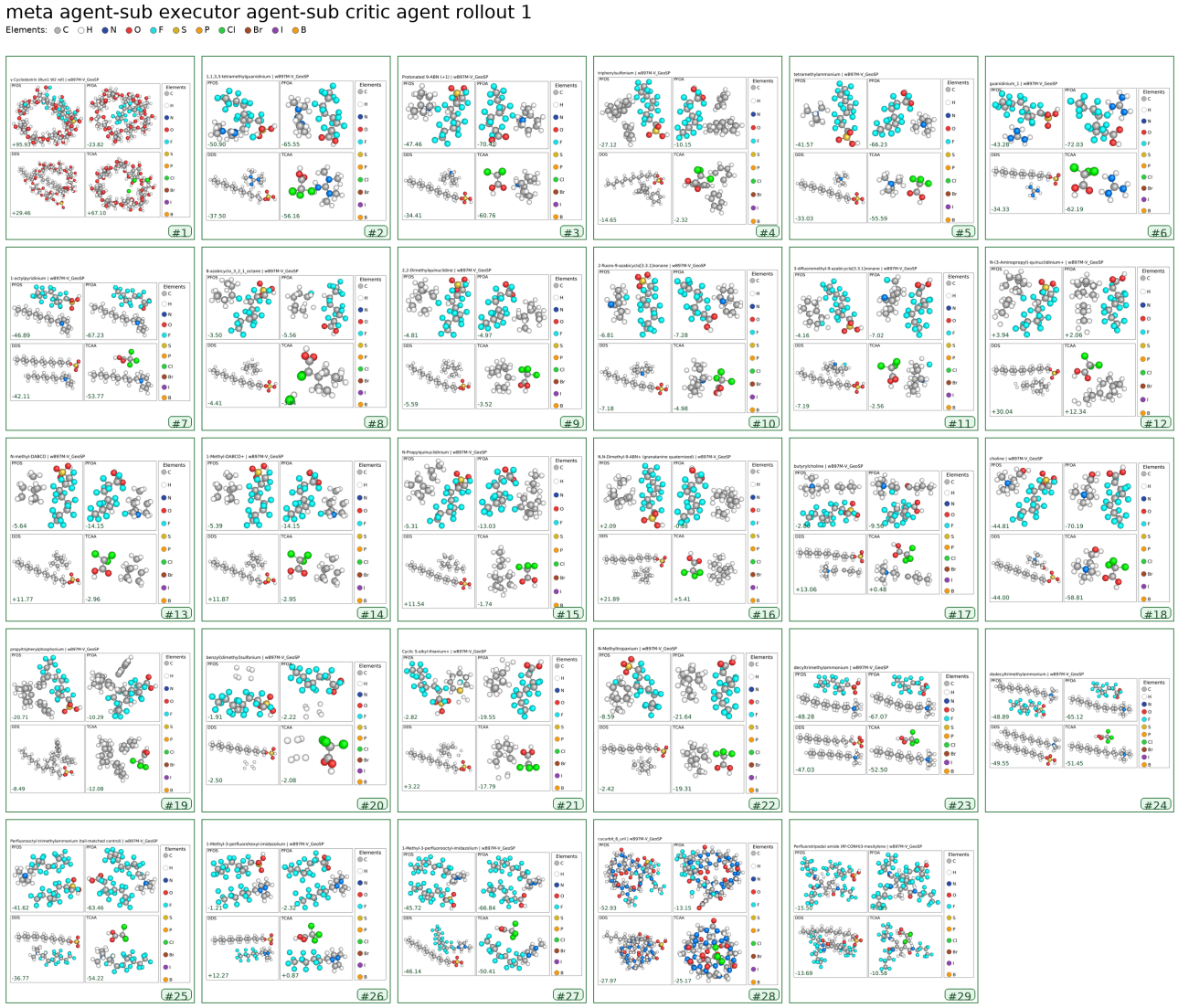


Figure S25. Meta-agent / sub-executor / sub-critic rollout 1.

Meta-agent / sub-executor-agent / sub-critic-agent rollouts (1–6).

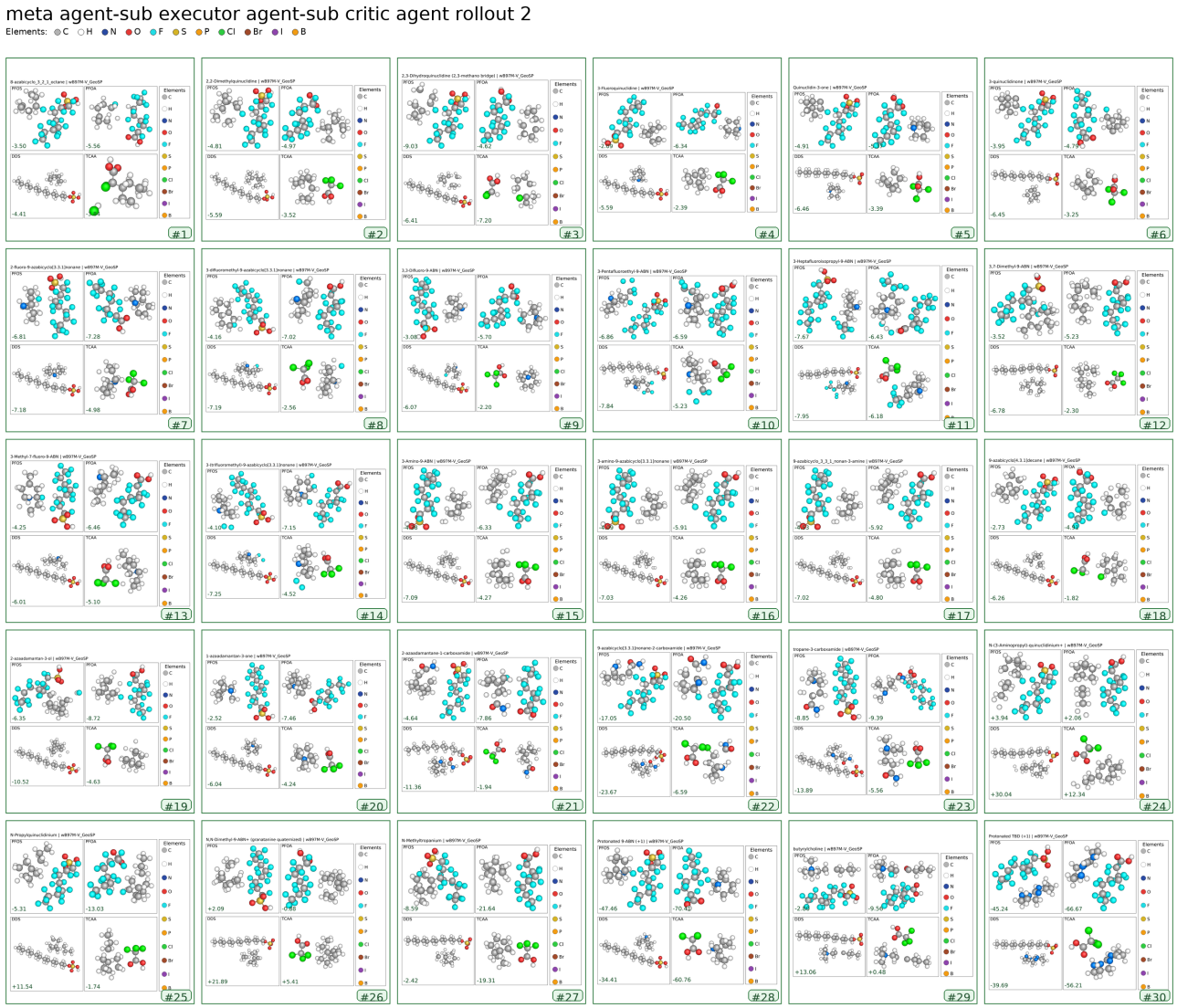


Figure S26. Meta-agent / sub-executor / sub-critic rollout 2.

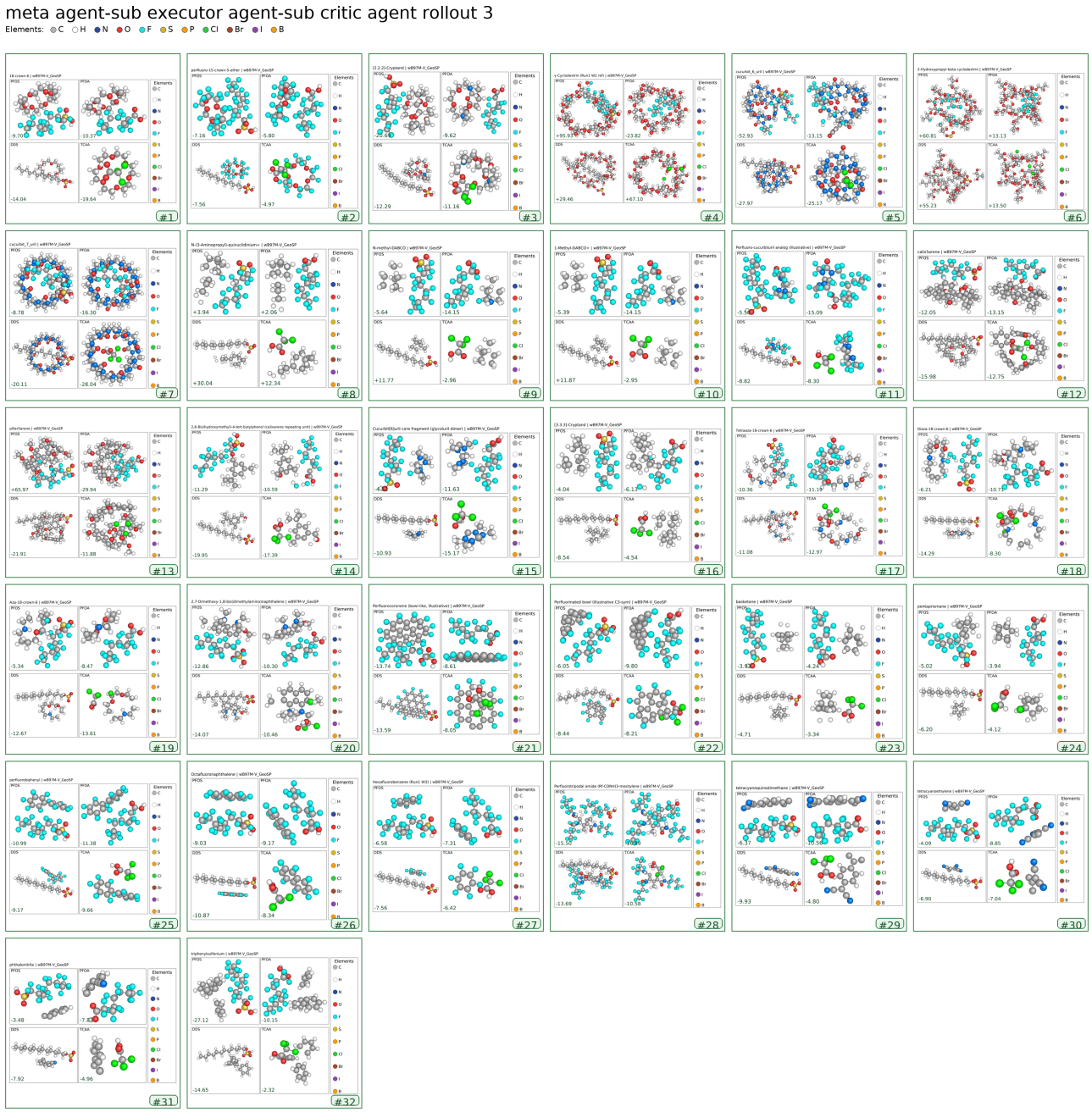


Figure S27. Meta-agent / sub-executor / sub-critic rollout 3.

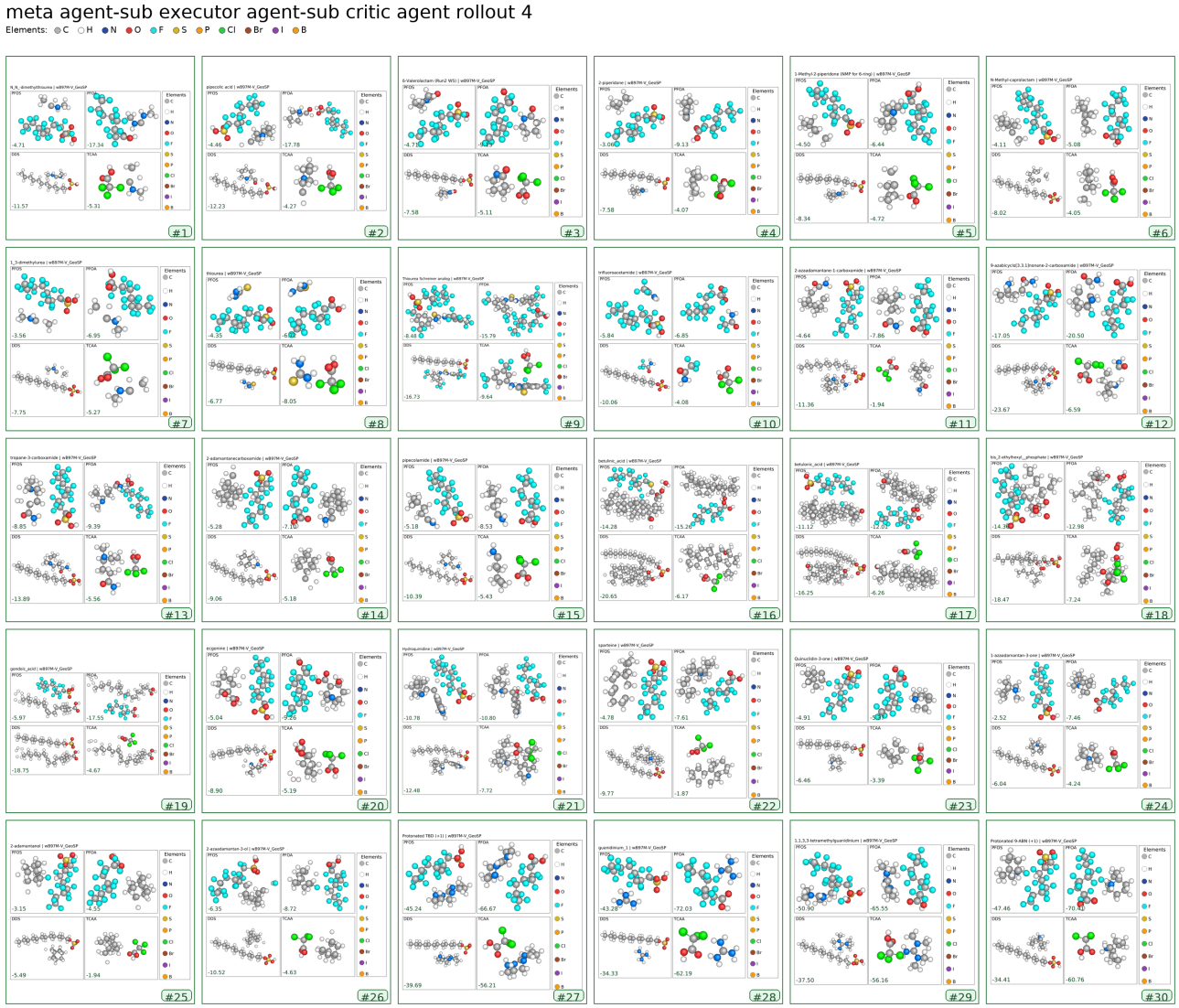


Figure S28. Meta-agent / sub-executor / sub-critic rollout 4.

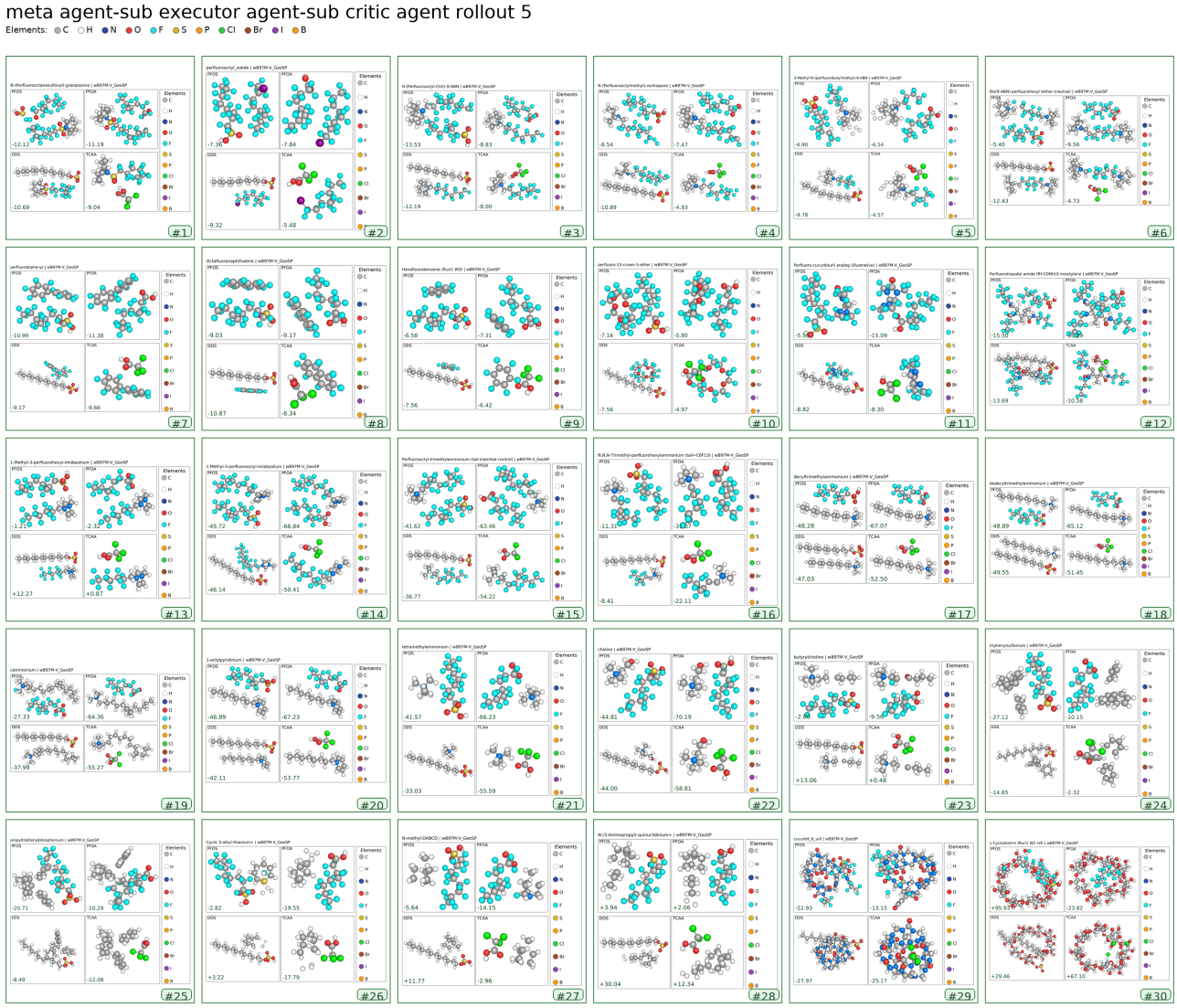


Figure S29. Meta-agent / sub-executor / sub-critic rollout 5.

Figure S30. Meta-agent / sub-executor / sub-critic rollout 6.

D. Supplementary tables

D.1. Replay-policy inventory

Table 3. Inventory of the 40 replay policies reported from the offline-replay study. Deployability labels follow Appendix B.5: *Direct* denotes no replay-time benchmark-reference updates, *Delayed-GT* requires per-system ground-truth feedback, *Retrospective* denotes selectors replayed with full-information or post hoc feedback, and *Oracle* is the unattainable lower bound. Neutral and charged aggregates are medians across the corresponding benchmark groups.

Strategy	Category	Deployability	Neutral MAE	Neutral ρ	Neutral h	Charged MAE	Charged ρ	Charged h
Always-CREST_xTB	Always-X	Direct	2.145	0.623	109.4	9.896	0.643	225.7
Always-CREST_xTB_DFT	Always-X	Direct	1.802	0.682	125.3	9.007	0.710	294.5
Always-PBE-D3BJ_GeoSP	Always-X	Direct	1.397	0.716	33.8	8.633	0.667	46.7
Always-PBE-D3BJ_SP	Always-X	Direct	1.504	0.613	12.3	12.780	0.561	22.9
Always-RAPIDS	Always-X	Direct	1.365	0.658	14.8	20.520	-0.043	20.4
Always-wB97M-V_GeoSP	Always-X	Direct	1.300	0.728	64.8	8.292	0.702	98.5
Always-wB97M-V_SP	Always-X	Direct	1.439	0.641	16.9	12.649	0.572	33.1
Always-wB97X-D3BJ_GeoSP	Always-X	Direct	1.437	0.728	77.4	8.574	0.694	116.8
Always-wB97X-D3BJ_SP	Always-X	Direct	1.418	0.630	15.0	12.549	0.574	27.2
Oracle	Oracle	Oracle	0.386	0.943	52.8	4.642	0.886	145.3
ALORS	Bandit	Retrospective	1.147	0.688	60.0	8.951	0.654	177.3
BiasCorr-RAPIDS	Bandit	Delayed-GT	1.304	0.621	14.8	20.067	-0.058	20.4
CheapEnsemble-3	Bandit	Direct	1.398	0.633	46.5	12.771	0.557	70.5
CheapEnsemble-5	Bandit	Direct	1.399	0.633	111.4	12.565	0.566	150.2
Cost-Aware	Bandit	Retrospective	1.172	0.680	76.4	10.589	0.571	110.2
Disagreement-t1	Bandit	Direct	1.379	0.654	35.6	8.597	0.703	262.4
Ladder-t1	Bandit	Direct	1.489	0.639	32.7	10.665	0.578	129.7
Learned-Selector	Bandit	Retrospective	1.054	0.714	90.6	8.057	0.688	275.6
MFMI-Greedy	Bandit	Retrospective	1.429	0.661	142.5	9.280	0.589	320.7
MFMI-Greedy-5x	Bandit	Retrospective	1.429	0.661	142.5	9.280	0.589	320.7
Random	Bandit	Direct	1.546	0.657	58.1	11.245	0.547	98.5
Stacking-r0	Bandit	Delayed-GT	1.158	0.615	51.3	9.152	0.648	151.3
StackingML-2	Bandit	Delayed-GT	1.384	0.596	30.5	9.339	0.507	43.3
Thompson	Bandit	Retrospective	1.292	0.683	50.3	8.882	0.648	177.9
Thompson-CostAware	Bandit	Retrospective	1.531	0.631	37.0	9.291	0.643	267.6
UCB	Bandit	Retrospective	1.042	0.700	61.7	8.335	0.648	211.3
UCB-CostAware	Bandit	Retrospective	1.194	0.714	86.5	9.549	0.648	284.7
BucketLearned	Chem/Bucket	Retrospective	1.085	0.707	102.1	8.202	0.679	233.1
BucketPrior-Adaptive	Chem/Bucket	Retrospective	1.024	0.756	55.5	8.376	0.659	188.5
ChemALORS	Chem/Bucket	Retrospective	1.167	0.680	62.7	9.312	0.613	137.5
ChemBiasCorr-PBE	Chem/Bucket	Delayed-GT	1.240	0.627	12.3	8.779	0.605	22.9
ChemLearned-Selector	Chem/Bucket	Retrospective	1.067	0.705	102.1	8.582	0.651	231.1
ChemStacking-3	Chem/Bucket	Delayed-GT	1.246	0.642	46.5	8.862	0.592	70.5
ChemUCB-8	Chem/Bucket	Retrospective	1.007	0.762	65.1	8.115	0.668	194.8
GBM-Stack-2cheap	GBM/Meta	Delayed-GT	1.100	0.708	30.5	8.467	0.505	43.3
GBM-Stack-MetaWarm	GBM/Meta	Delayed-GT	1.319	0.707	30.5	9.415	0.545	43.3
Hybrid-SelectStack	GBM/Meta	Delayed-GT	1.114	0.734	131.9	9.381	0.589	321.5
LinUCB	GBM/Meta	Retrospective	1.617	0.649	74.5	8.984	0.635	188.9
Meta-GBM	GBM/Meta	Direct	1.551	0.574	30.5	14.091	0.288	43.3
Meta-GBM-Adaptive	GBM/Meta	Delayed-GT	1.346	0.694	30.5	10.405	0.539	43.3

D.2. Full 18 × 9 per-benchmark MAE matrix

Table 4 and Table 5 give the per-benchmark capped MAE (kcal/mol) of every fidelity method on every benchmark, where each method's relaxed binding score is compared against the published benchmark interaction-energy reference. These are the numbers summarized by the full appendix heatmap in Fig. S1.

Table 4. Per-benchmark capped MAE (kcal/mol), part A (benchmarks 1–9), computed between each method's relaxed binding score and the published benchmark interaction-energy reference.

Method	A24	S66	HB300SPX	HB375	X40	D1200_Hal.	SH250	D1200_HBCNO	D1200_PS
RAPIDS	1.68	1.60	1.43	1.74	1.30	0.59	2.23	0.87	0.95
PBE-D3BJ SP	2.05	1.63	1.38	1.64	1.23	0.62	2.11	0.51	0.69
ωB97X-D3BJ SP	1.76	2.05	1.22	1.62	1.00	0.59	2.09	0.57	0.70
ωB97M-V SP	1.78	1.61	1.27	1.64	0.99	0.57	2.18	0.64	0.84
PBE-D3BJ GeoSP	0.76	1.74	1.70	1.35	1.14	0.82	2.23	0.36	0.73
ωB97X-D3BJ GeoSP	1.79	3.19	0.91	1.36	0.93	0.59	1.81	0.36	0.57
ωB97M-V GeoSP	1.43	1.44	0.89	1.28	0.95	0.49	1.89	0.37	0.48
CREST xTB	1.96	1.06	2.11	3.01	5.14	1.40	4.08	0.90	2.18
CREST xTB+DFT	1.52	0.72	1.17	2.37	1.43	0.95	1.85	1.04	2.23

Table 5. Per-benchmark capped MAE (kcal/mol), part B (benchmarks 10–18), computed between each method's relaxed binding score and the published benchmark interaction-energy reference.

Method	BBI	HSG	NBC1	SSI-disp.	SSI-mixed	SSI-other	SSI-polar	IHB100	SSI-charged
RAPIDS	3.29	3.69	0.53	0.86	1.25	0.78	3.79	13.83	27.21
PBE-D3BJ SP	3.73	4.61	0.44	1.14	1.80	1.02	4.69	8.03	17.53
ωB97X-D3BJ SP	3.41	4.90	0.56	1.09	1.91	0.95	4.41	7.81	17.29
ωB97M-V SP	3.47	4.85	0.56	0.96	1.81	0.83	4.45	7.80	17.50
PBE-D3BJ GeoSP	2.99	3.80	0.51	1.45	2.32	1.29	5.20	5.04	12.23
ωB97X-D3BJ GeoSP	2.93	3.60	0.62	1.52	2.27	1.27	4.82	5.63	11.52
ωB97M-V GeoSP	3.02	3.59	0.50	1.32	2.17	1.13	4.82	4.90	11.68
CREST xTB	2.39	3.17	1.32	1.65	2.62	1.25	6.10	9.23	10.57
CREST xTB+DFT	2.63	4.41	1.29	2.22	3.04	1.76	6.82	5.79	12.22

D.3. Per-benchmark MAE ladder for the decomposition of §4.3

Table 6 gives the per-benchmark values of the benchmark-relative MAE ladder used in §4.3. This ladder tracks how the benchmark-reference MAE changes when one rung of the fidelity stack is swapped for the next, but it is not itself a direct estimate of $\mathrm{MAE}(|\delta_{\mathrm{surf}}|)$, $\mathrm{MAE}(|\delta_{\mathrm{geom}}|)$, or $\mathrm{MAE}(|\delta_{\mathrm{samp}}|)$ from the system-level decomposition. To avoid colliding with the system-level decomposition symbols $\delta_{\mathrm{surf}}(s)$, $\delta_{\mathrm{geom}}(s)$, and $\delta_{\mathrm{samp}}(s)$ used in the main text, the table reports the corresponding *adjacent MAE differences* as $\Delta\mathrm{MAE}_{\mathrm{surf}}$, $\Delta\mathrm{MAE}_{\mathrm{geom}}$, and $\Delta\mathrm{MAE}_{\mathrm{samp}}$. Entries are capped MAE (kcal/mol) of the four ladder rungs and their pairwise differences.

D.4. Per-benchmark validation-guard trigger rates

D.5. Per-benchmark guard flagged-vs-unflagged capped-MAE

Table 8 reports the per-benchmark capped-MAE of the RAPIDS predictions broken out by whether the validation stack (§3.2) emits a flagged label. These are the numbers summarized at the aggregate level in the main-text reliability note at the end of §4.5.

Table 6. Per-benchmark benchmark-relative MAE ladder and adjacent MAE differences. Columns: capped MAE of RAPIDS, ωB97M-V/def2-TZVP SP on RAPIDS geometry, ωB97M-V GeoSP on RAPIDS geometry, and CREST xTB+DFT, each compared against the published benchmark interaction-energy reference; then the three adjacent differences $\Delta\mathrm{MAE}_{\mathrm{surf}} = \mathrm{MAE}_{\mathrm{RAPIDS}} - \mathrm{MAE}_{\mathrm{wB97M\text{-}V_SP}}$, $\Delta\mathrm{MAE}_{\mathrm{geom}} = \mathrm{MAE}_{\mathrm{wB97M\text{-}V_SP}} - \mathrm{MAE}_{\mathrm{wB97M\text{-}V_GeoSP}}$, $\Delta\mathrm{MAE}_{\mathrm{samp}} = \mathrm{MAE}_{\mathrm{wB97M\text{-}V_GeoSP}} - \mathrm{MAE}_{\mathrm{CREST_xTB_DFT}}$. Negative $\Delta\mathrm{MAE}$ indicates the higher-fidelity rung has a *worse* MAE against the benchmark reference, which is the benchmark-relative pattern identified on the BFDb SSI neutral subsets (§4.2).

Benchmark	RAPIDS	ωB97M-V/def2-TZVP SP	ωB97M-V GeoSP	CREST xTB+DFT	$\Delta\mathrm{MAE}_{\mathrm{surf}}$	$\Delta\mathrm{MAE}_{\mathrm{geom}}$	$\Delta\mathrm{MAE}_{\mathrm{samp}}$
A24	1.68	1.78	1.43	1.52	-0.10	$+0.35$	-0.09
S66	1.60	1.61	1.44	0.72	-0.01	$+0.17$	$+0.72$
HB300SPX	1.43	1.27	0.89	1.17	$+0.16$	$+0.37$	-0.28
HB375	1.74	1.64	1.28	2.37	$+0.10$	$+0.36$	-1.09
X40	1.30	0.99	0.95	1.43	$+0.31$	$+0.04$	-0.48
D1200_Halogens	0.59	0.57	0.49	0.95	$+0.02$	$+0.08$	-0.46
SH250	2.23	2.18	1.89	1.85	$+0.04$	$+0.29$	$+0.04$
D1200_HBCNO	0.87	0.64	0.37	1.04	$+0.24$	$+0.26$	-0.67
D1200_PS	0.95	0.84	0.48	2.23	$+0.10$	$+0.37$	-1.76
BFDb_BBI	3.29	3.47	3.02	2.63	-0.19	$+0.45$	$+0.39$
BFDb_HSG	3.69	4.85	3.59	4.41	-1.16	$+1.26$	-0.82
BFDb_NBC1	0.53	0.56	0.50	1.29	-0.04	$+0.06$	-0.79
BFDb_SSI_dispersion	0.86	0.96	1.32	2.22	-0.09	-0.36	-0.90
BFDb_SSI_mixed	1.25	1.81	2.17	3.04	-0.56	-0.35	-0.87
BFDb_SSI_other	0.78	0.83	1.13	1.76	-0.05	-0.30	-0.62
BFDb_SSI_polar	3.79	4.45	4.82	6.82	-0.65	-0.38	-2.00
IHB100	13.83	7.80	4.90	5.79	$+6.03$	$+2.90$	-0.89
BFDb_SSI_charged	27.21	17.50	11.68	12.22	$+9.72$	$+5.81$	-0.54

Table 7. RAPIDS validation-guard trigger rates per benchmark. A system is counted as flagged when at least one element of the validation stack downgrades the confidence from high, namely one of the four per-configuration guards (topology / geometry / energy / convergence) or the scan-level Energy Consistency Guard (ECG). N_{missing} counts systems with no committed RAPIDS prediction because at least one required lower-level engine in the benchmarked pipelines (UMA or CREST/xTB preprocessing) does not support the chemistry. The trigger rate is computed on completed predictions, i.e. $flagged/N_{\mathrm{completed}}$. Data also recorded in benchmark_datasets.md.

Benchmark	$N_{\mathrm{completed}}$	N_{missing}	**flagged**	**rate**
A24	22	2	5	22.7%
S66	66	0	0	0.0%
X40	40	0	4	10.0%
HB300SPX	300	0	7	2.3%
HB375	375	0	0	0.0%
SH250	250	0	23	9.2%
D1200_HBCNO	308	0	10	3.2%
D1200_PS	293	0	56	19.1%
D1200_Halogens	310	0	10	3.2%
BFDb_BBI	100	0	17	17.0%
BFDb_HSG	21	0	4	19.0%
BFDb_NBC1	10	0	2	20.0%
BFDb_SSI_dispersion	1,325	0	218	16.5%
BFDb_SSI_mixed	765	0	129	16.9%
BFDb_SSI_other	240	0	33	13.8%
BFDb_SSI_polar	266	0	47	17.7%
IHB100	100	0	4	4.0%
BFDb_SSI_charged	776	0	429	55.3%
Total	**5,567**	**2**	**998**	**17.9%**

Table 8. RAPIDS capped MAE on the unflagged (ok) subset versus the flagged subset, with an explicit missing count for systems that yield no committed RAPIDS prediction because at least one required lower-level engine (UMA or CREST/xTB preprocessing in the benchmarked pipelines) does not support the chemistry. Benchmarks with zero flagged predictions (S66, HB375) have no ratio defined; ratio < 1 indicates the flagged subset carries a *lower* MAE than the unflagged subset, which we observe on two benchmarks with small n_{flagged}.

Benchmark	N_{ok}	N_{flagged}	N_{missing}	$\mathrm{MAE}_{\mathrm{ok}}$	$\mathrm{MAE}_{\mathrm{flagged}}$	ratio
A24	17	5	2	0.77	4.78	6.22
S66	66	0	0	1.60	—	—
X40	36	4	0	1.00	4.01	4.01
HB300SPX	293	7	0	1.42	2.01	1.42
HB375	375	0	0	1.74	—	—
SH250	227	23	0	2.01	4.33	2.15
D1200_HBCNO	298	10	0	0.86	1.19	1.38
D1200_PS	237	56	0	0.93	1.01	1.09
D1200_Halogens	300	10	0	0.58	0.89	1.52
BFDb_BBI	83	17	0	3.08	4.30	1.40
BFDb_HSG	17	4	0	3.69	3.66	0.99
BFDb_NBC1	8	2	0	0.43	0.94	2.19
BFDb_SSI_dispersion	1107	218	0	0.83	1.04	1.25
BFDb_SSI_mixed	636	129	0	1.16	1.71	1.47
BFDb_SSI_other	207	33	0	0.75	0.96	1.28
BFDb_SSI_polar	219	47	0	3.93	3.14	0.80
IHB100	96	4	0	14.08	7.79	0.55
BFDb_SSI_charged	347	429	0	26.54	27.76	1.05
Total (all 18 bm)	4569	998	2	3.48	12.91	3.71