Hierarchical Deep Research with Local–Web RAG: Toward Automated System-Level Materials Discovery

Technical Appendices and Supplementary Material

A Supplementary Figures

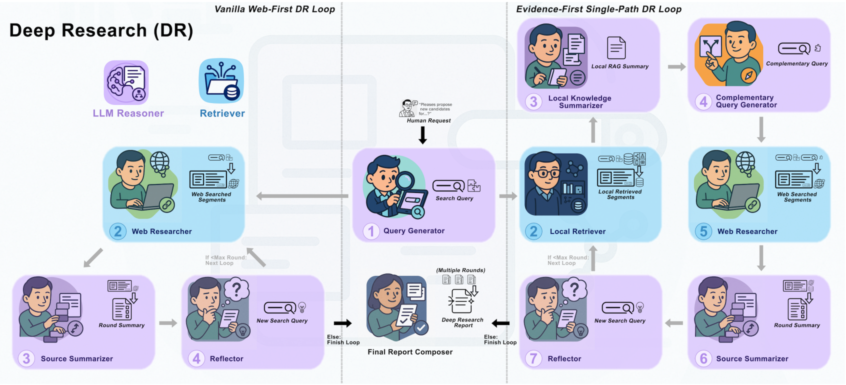


Figure S1: Schematic of the local evidence-first single DR instance in comparison with common web-first DR (Adapted from [22])

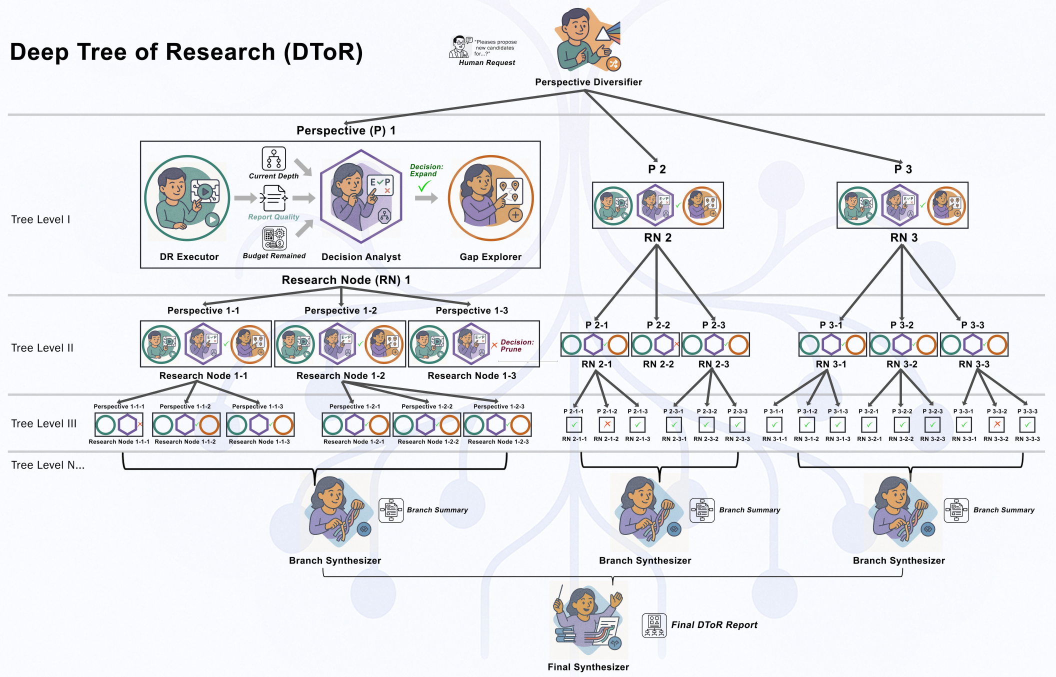


Figure S2: Schematic of the DToR workflow

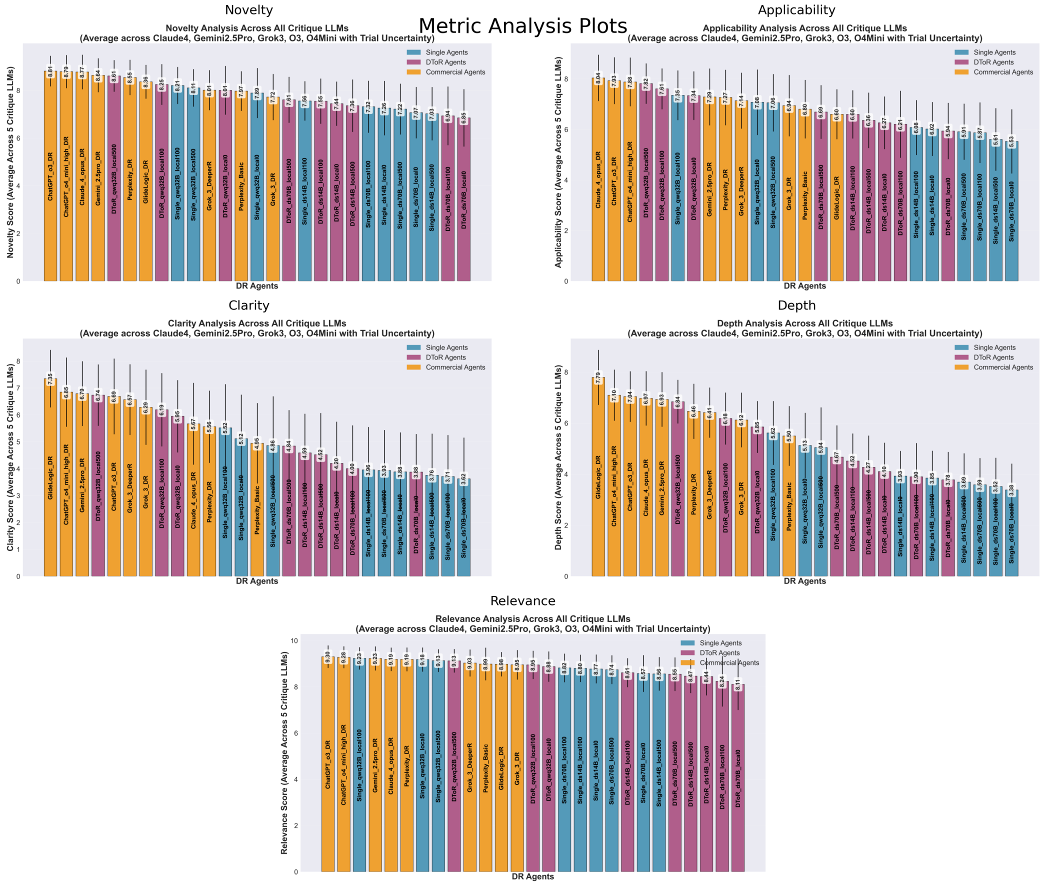


Figure S3: Overall rubric scores across 27 agents on different evaluation dimensions.

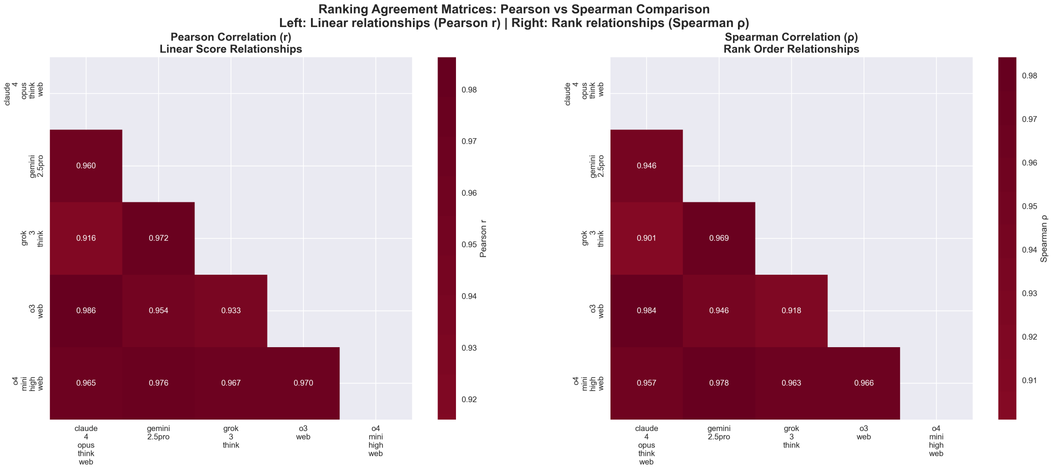
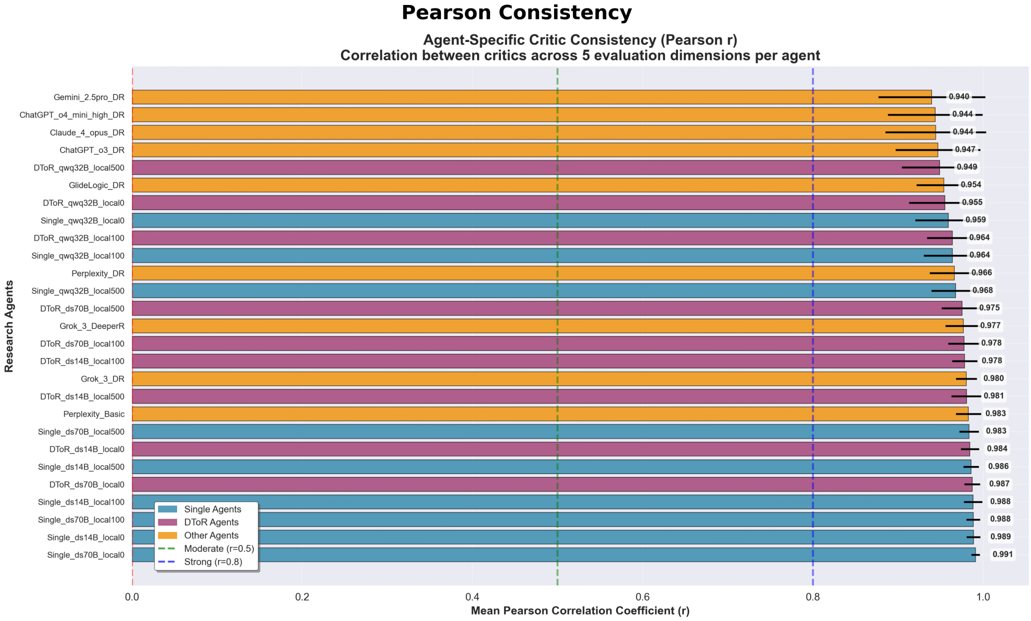


Figure S4: Rubric score agent agreement level over the five critic LLMs, computed by correlation factors.



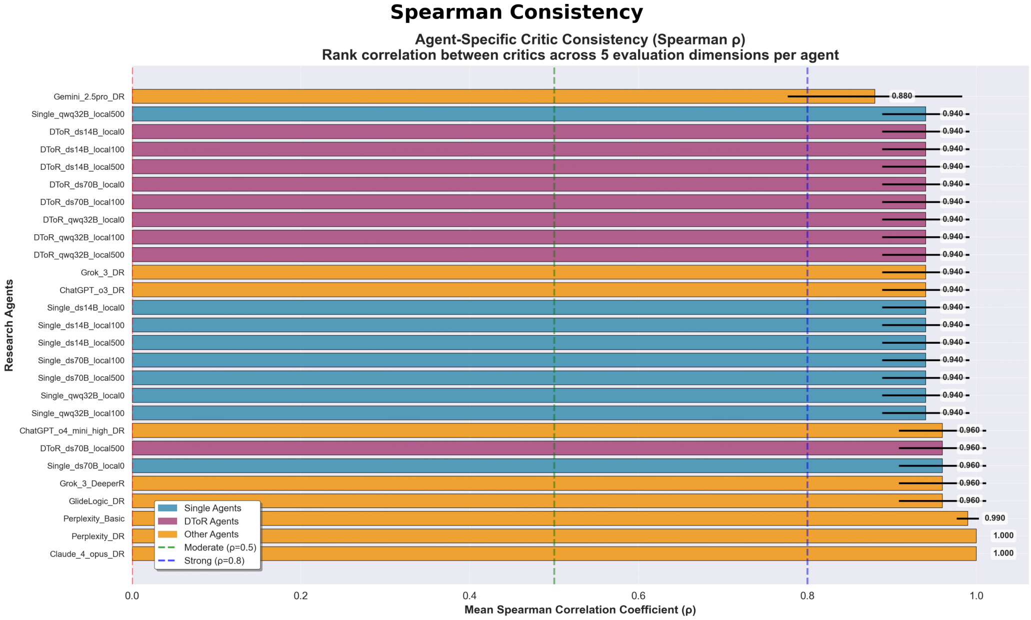


Figure S5: Agent-Specific Critic Consistency Analysis. Pearson correlation coefficients measuring linear agreement between critique LLMs across 5 evaluation dimensions (Relevance, Depth, Clarity, Applicability, Novelty) for each research agent. Spearman rank correlation coefficients measuring ordinal agreement between critics. Higher values indicate greater consensus among the 5 critique LLMs (Claude, Gemini, Grok, O3, O4) about each agent's relative performance strengths and weaknesses. Data aggregated from 21 research topics × 3 trials per agent. Perfect correlation ($\rho = 1.00$) indicates all critics completely agree on the ranking of an agent's performance across the 5 dimensions.

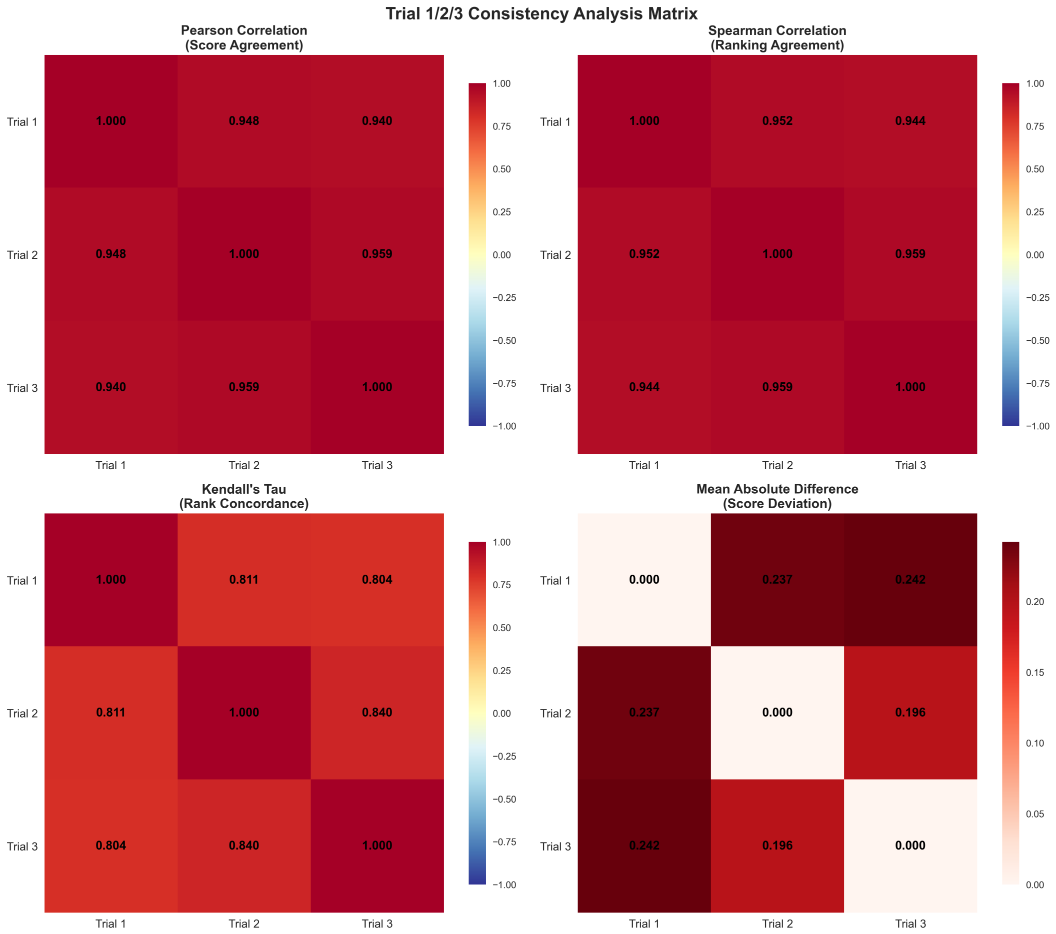
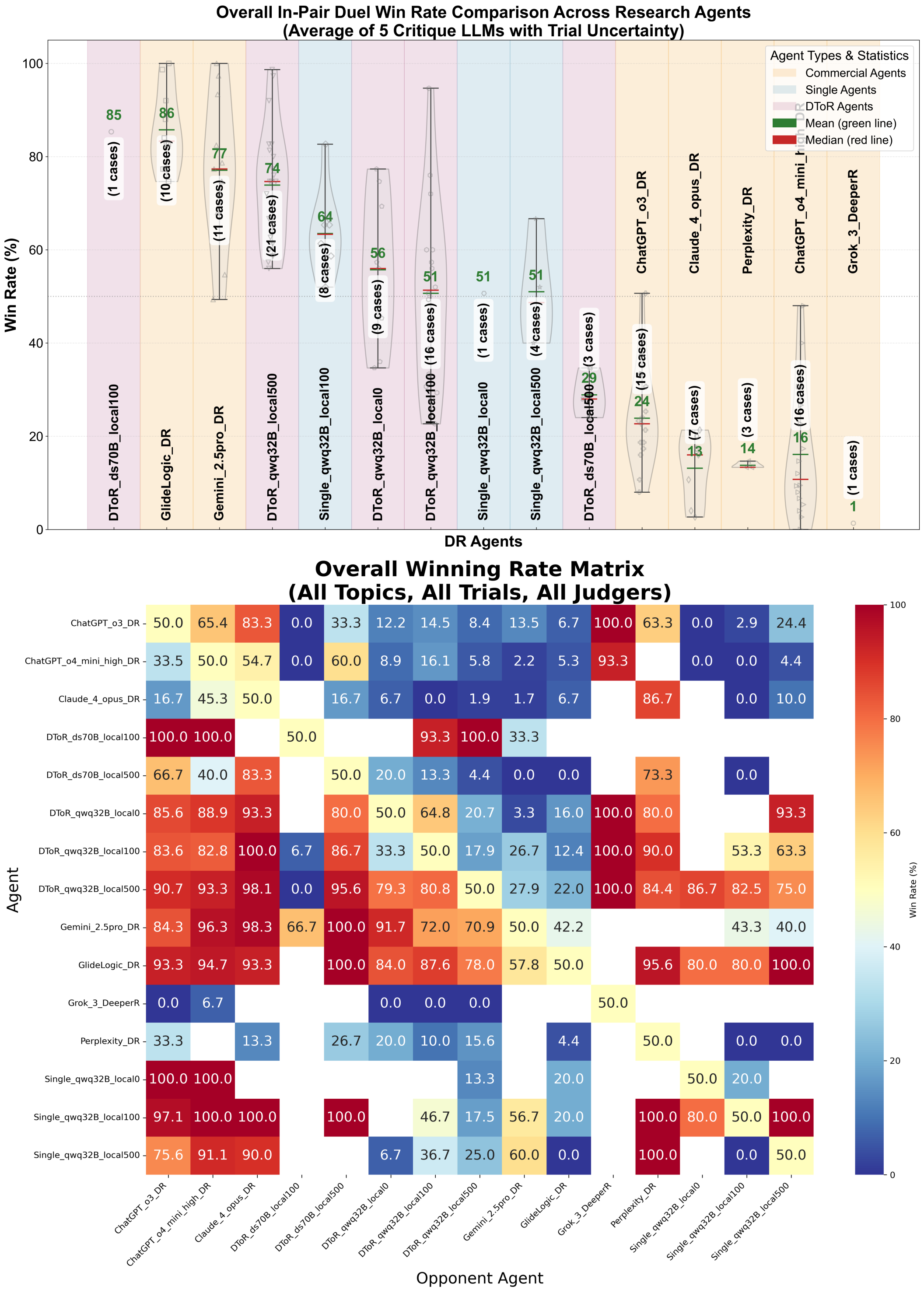
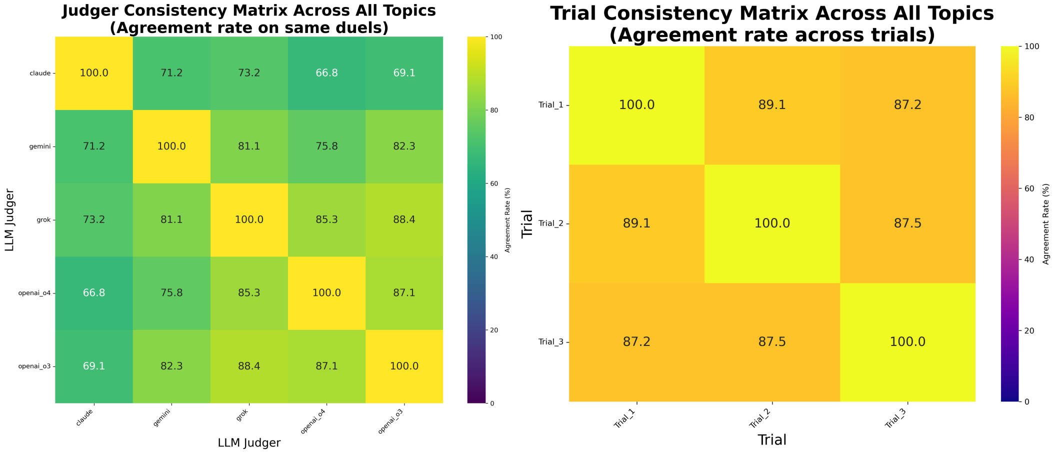


Figure S6: Trial-to-Trial Consistency Analysis Matrix. Four metrics assess reproducibility across three independent trials: (A) Pearson correlation measures score agreement, (B) Spearman correlation evaluates ranking consistency, (C) Kendall's tau provides alternative rank concordance, and (D) Mean absolute difference quantifies score deviations. Data aggregated across 21 topics, 27 agents, and 5 LLM critics. Higher correlations (>0.8) and lower deviations (<0.5) indicate excellent experimental reproducibility.



Agent	ChatGPT_o3_DR	ChatGPT_o4_mini_high_DR	Claude_4_opus_DR	DToR_ds70B_local100	DToR_ds70B_local500	DToR_qwq32B_local0	DToR_qwq32B_local100	DToR_qwq32B_local500	Gemini_2.5pro_DR	GlideLogic_DR	Grok_3_DeeperR	Perplexity_DR	Single_qwq32B_local0	Single_qwq32B_local100	Single_qwq32B_local500
ChatGPT_o3_DR	50.0	65.4	83.3	0.0	33.3	12.2	14.5	8.4	13.5	6.7	100.0	63.3	0.0	2.9	24.4
ChatGPT_o4_mini_high_DR	33.5	50.0	54.7	0.0	60.0	8.9	16.1	5.8	2.2	5.3	93.3		0.0	0.0	4.4
Claude_4_opus_DR	16.7	45.3	50.0		16.7	6.7	0.0	1.9	1.7	6.7		86.7		0.0	10.0
DToR_ds70B_local100	100.0	100.0		50.0			93.3	100.0	33.3						
DToR_ds70B_local500	66.7	40.0	83.3		50.0	20.0	13.3	4.4	0.0	0.0		73.3		0.0	
DToR_qwq32B_local0	85.6	88.9	93.3		80.0	50.0	64.8	20.7	3.3	16.0	100.0	80.0			93.3
DToR_qwq32B_local100	83.6	82.8	100.0	6.7	86.7	33.3	50.0	17.9	26.7	12.4	100.0	90.0		53.3	63.3
DToR_qwq32B_local500	90.7	93.3	98.1	0.0	95.6	79.3	80.8	50.0	27.9	22.0	100.0	84.4	86.7	82.5	75.0
Gemini_2.5pro_DR	84.3	96.3	98.3	66.7	100.0	91.7	72.0	70.9	50.0	42.2				43.3	40.0
GlideLogic_DR	93.3	94.7	93.3		100.0	84.0	87.6	78.0	57.8	50.0		95.6	80.0	80.0	100.0
Grok_3_DeeperR	0.0	6.7				0.0	0.0	0.0			50.0				
Perplexity_DR	33.3		13.3		26.7	20.0	10.0	15.6		4.4		50.0		0.0	0.0
Single_qwq32B_local0	100.0	100.0						13.3		20.0			50.0	20.0	
Single_qwq32B_local100	97.1	100.0	100.0		100.0		46.7	17.5	56.7	20.0		100.0	80.0	50.0	100.0
Single_qwq32B_local500	75.6	91.1	90.0			6.7	36.7	25.0	60.0	0.0		100.0		0.0	50.0

Figure S7: Upper: Win Rate Violin Plot - Shows the distribution of win rates across different agents and topics with detailed statistical information and performance variations. Bottom: Overall Winning Matrix All Topics - Displays the comprehensive winning matrix with head-to-head performance comparisons between all agents across all research topics.


Judger Consistency Matrix Across All Topics (Agreement rate on same duels)

LLM Judger	claude	gemini	grok	openai_o4	openai_o3
claude	100.0	71.2	73.2	66.8	69.1
gemini	71.2	100.0	81.1	75.8	82.3
grok	73.2	81.1	100.0	85.3	88.4
openai_o4	66.8	75.8	85.3	100.0	87.1
openai_o3	69.1	82.3	88.4	87.1	100.0

Trial Consistency Matrix Across All Topics (Agreement rate across trials)

Trial	Trial_1	Trial_2	Trial_3
Trial_1	100.0	89.1	87.2
Trial_2	89.1	100.0	87.5
Trial_3	87.2	87.5	100.0

Figure S8: Consistency of in pair dueling. Left: Judger Consistency All Topics - Shows the consistency of judgment patterns across different research topics and agents, indicating how reliable the evaluation criteria are. Right: Trial Consistency All Topics - Displays the consistency of performance across multiple trials for each agent-topic combination, revealing the stability and reproducibility of results.

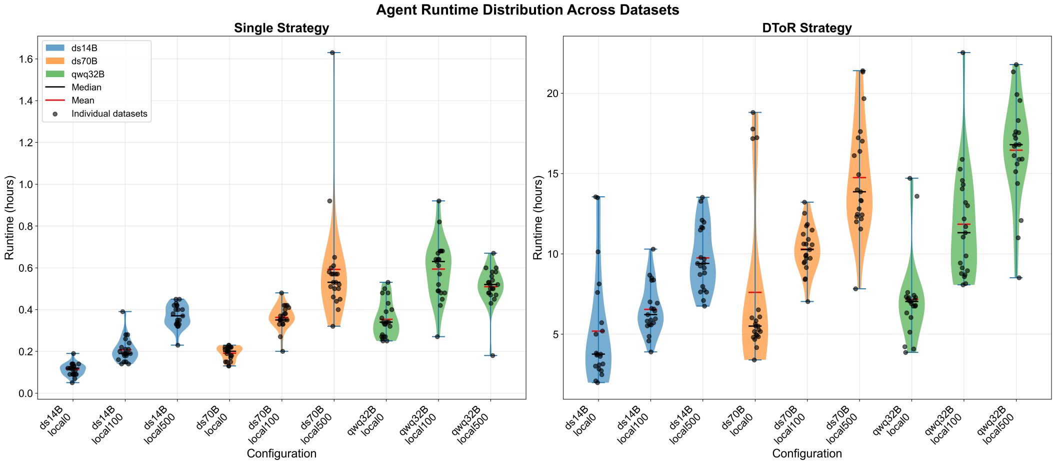


Figure S9: Violin plot visualizing the distribution of local DR agents runtime under different settings.

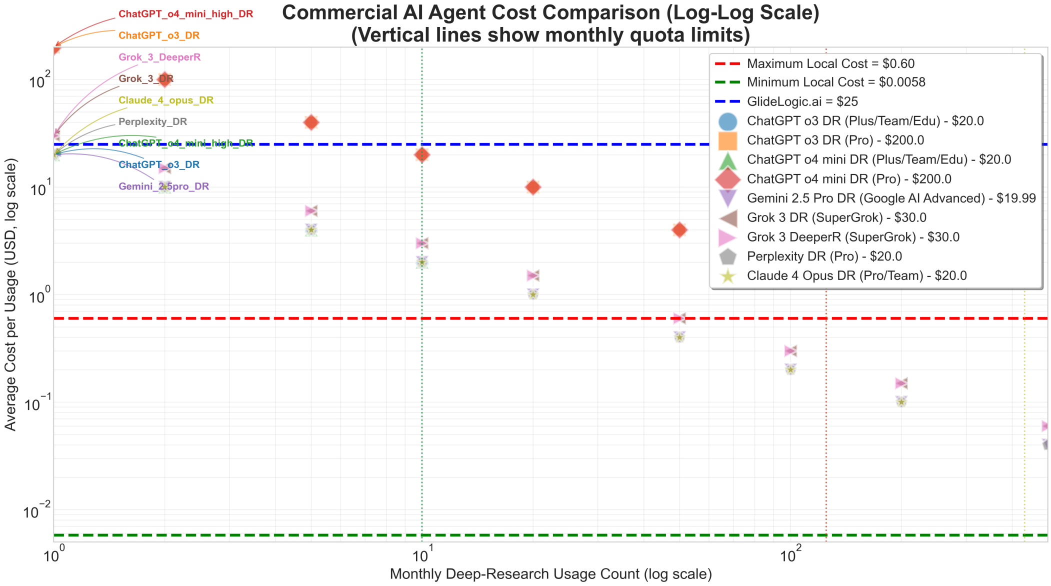


Figure S10: Comparison of cost over commercial DR agent solution with our local solution based on Chicago area electricity price.

B Related Work in Physics ML Surrogates, Domain LLMs, and Agentic Systems

Physics-aligned surrogates and classical ML. Physics-guided predictive surrogates now operate at web-scale, with graph networks and equivariant models trained on massive curated corpora. For crystalline stability, GNoME trains large graph models and active-learning pipelines to propose millions of candidate structures and flag hundreds of thousands as likely stable—establishing that data+scale can systematically widen the known inorganic universe [6]. In catalysis, the Open Catalyst efforts formalized realistic workloads: OC20 introduced 1.28M DFT relaxations and 265M single-point evaluations for adsorbate–surface energetics and created standardized tasks (IS2RE/IS2RS/S2EF) with stringent generalization splits; OC22 extended to oxide electrocatalysts relevant to OER with more realistic catalysts and targets [7, 24]. New foundation-style datasets push further: OMat24 aggregates >110M DFT entries and releases strong baselines (e.g., Equiformer-V2) that achieve near-chemical-accuracy on multiple targets [25]; OMol25 analogously scales molecular DFT data to >100M calculations spanning tens of millions of unique molecules, enabling broad pretraining for small-molecule physics [8]. Complementing datasets, universal model of atoms (UMA) explore single architectures that transfer across chemistry and tasks, aiming for cross-domain generalization with shared tokenization/featurization of atoms and local environments [26]. These surrogates excel at S1–S2 targets (bulk stability, adsorption energies, local structural relaxations) but are largely D1–D2 in operation: fixed-form predictors trained on equilibrium data, with limited capacity to reason about kinetics, synthesis windows, or device-level constraints that emerge at S3–S4. The net effect is that they remain superb "atoms-to-properties" engines, yet they do not autonomously help solve the long-horizon, multi-paper evidence assembly and process-aware planning needed, from atomic scale to device scale rationales [6–8, 24–26].

Domain LLMs (non-agent). Domain LLMs span three useful roles. (i) Materials-text models (MatSciBERT/MatBERT/MaterialBERT) are trained on materials corpora to power entity/relation extraction, document classification, and graph building; MatSciBERT outperforms SciBERT on named entity recognition (NER) and relation tasks across curated benchmarks, while MatBERT shows that domain-specific pretraining significantly lifts NER F1 and sample efficiency in materials texts; MaterialBERT introduces a domain tokenizer/vocabulary tuned to materials literature [27–29]. (ii) Chemical/polymer LLMs target sequence-level prediction/generation: ChemBERTa established transformer baselines and scaling trends for SMILES; ChemBERTa-2 pushed to 77M SMILES and showed consistent gains (and that multi-task regression pretraining can beat (Masked Language Modeling) MLM in transfer); ChemBERTa-3 contributes open, end-to-end training infrastructure for chemical foundation models; MegaMolBART provides a large-scale generative sequence model optimized for molecular design [30–33]. ChemGPT demonstrates generative scaling to >1B parameters with clear neural-scaling laws for chemical LLMs, clarifying how model/data scale trade off in practice [34]. In polymers, TransPolymer introduces a chemically-aware tokenizer and MLM pretraining that improves regression performances across ten benchmarks; polyBERT builds an end-to-end, ultrafast pipeline and screens 100M hypothetical polymers; PolyNC fuses natural-language and chemical-sequence inputs to improve unified polymer property prediction [35–37]. (iii) Generative/design LLMs for materials include ChatMOF, which conditions MOF proposals/properties from text and structural descriptors; Crystal-LLM (autoregressive) generates crystal structures from composition; and MatterGen scales inorganic structure generation with validated discoveries and robust property control [38–40]. Collectively these models deliver flexible text/code/structure interfaces and strong S1–S2 performance—ideal for curation, candidate suggestion, or surrogate-assisted screening—but they are typically short-horizon (D1–D2): they answer prompts, not conduct hours-long, tool-grounded, hierarchical inquiry across conflicting literature [27–40].

Agentic systems for scientific inquiry. Agent frameworks aim to orchestrate end-to-end inquiry: decompose goals, retrieve and reconcile evidence, use domain tools/simulators, and iteratively critique/plan. Early exemplars used rule/graph planning and Bayesian loops within constrained scopes—Chematica/SYNTHIA operationalized expert retrosynthesis rules and heuristic search; CAMD framed autonomous design/make/test/learn for materials screening [5, 41]. LLM-based agents extend this with language-native decomposition and tool-use: ChemCrow wires models to chemistry tools (e.g., RDKit, databases) to complete realistic tasks; HoneyComb provides a flexible materials-agent stack with retrieval/tooling and HPC hooks; SciAgents formalizes multi-agent graph reasoning for scientific planning; "AI co-scientist" frameworks couple planning with web/code tools and emphasize evaluation and provenance [42–45]. Execution back-ends close the loop: A-Lab demonstrates a self-driving lab discovering and characterizing new materials; IBM RXN/RoboRXN links digital retrosynthesis with automated synthesis; NVIDIA ALCHEMI targets scaled infrastructure (data, models, and accelerated simulation) for materials discovery [46–48]. Despite progress, most agents today operate at D2–D3 depth and S1–S3 scope; they rarely sustain thousand-query retrieval runs, explicit topic-tree governance, or tight coupling to cross-domain process constraints that S3–S4 device problems demand. This is precisely the niche for deep-research agents: to deliver hierarchical planning, multi-stage local+web retrieval at scale, targeted tool/simulator calls, and systematic self-critique with dense provenance until a decision-ready synthesis emerges [5,41–48].

Gap to deep research at S3–S4/D4. Despite existing progress, the transition to S3–S4 nano-device complexity poses a fundamental barrier. The combinatorial explosion across materials, interfaces, processing parameters, and operational constraints creates hypothesis spaces where decisive insights lie at the intersections of disparate literature, conflicting reports, and cross-domain phenomena. Crucially, long-horizon, hierarchical evidence accumulation (D4) remains under-documented: many agents cap at D2–D3, with limited iterative loops, shallow coordination, and no explicit topic-tree control for breadth-then-depth synthesis. Commercial deep-research offerings introduced in 2025 enable multi-step inquiry but remain closed-source with limited controllability and local integration, reinforcing the need for open, on-prem frameworks tailored to proprietary materials workflows.

C Pseudocode

Algorithm 1: Vanilla Web-First DR Agent(Adapted from [22])

```
Require: Inputs: research_topic $\in \mathbb{S}$; configuration $C = \{$ llm_provider$\in \{$lmstudio, ollama$\}$,
         search_api$\in \{$tavily, perplexity, duckduckgo, searxng$\}$, fetch_full_page$\in \{0, 1\}$,
         strip_thinking_tokens$\in \{0, 1\}$, use_tool_calling$\in \{0, 1\}$, max_web_research_loops$\in \mathbb{N}\}$.
Ensure: final running_summary (with deduplicated sources appended).
State: running_summary $\leftarrow$ ""; search_query $\leftarrow$ ""; web_research_results $\leftarrow$ [ ];
    sources_gathered $\leftarrow$ [ ]; research_loop_count $\leftarrow 1$.
function GenerateQuery(state, $C$)
    Build system prompt from research_topic and current_date.
    if $C$.use_tool_calling $= 1$ then
        Ask LLM to return tool payload $\{$query, rationale$\}$.
    else
        Ask LLM in JSON mode to return $\{$query, rationale$\}$.
    Parse JSON/tool output; let $q$ be the extracted query.
    if $q$ is empty or parsing fails then
        Strip $\langle$think$\rangle$ blocks if $C$.strip_thinking_tokens $= 1$.
        Fallback to first nonempty line; if still empty, set $q \leftarrow$"Tell me more about
            research_topic".
    search_query $\leftarrow q$; return $q$.
function WebResearch(state, $C$)
    Select backend $b \leftarrow C$.search_api.
    Issue web search with $q =$ search_query via $b$; obtain raw results $\mathcal{R}$.
    Format bullets $S \leftarrow$ format_sources($\mathcal{R}$).
    Build long string $W \leftarrow$ deduplicate_and_format_sources($\mathcal{R}, m, C$.fetch_full_page),
        where $m$ is the max tokens per source (vanilla uses $\approx 1000$).
    Append $S$ to sources_gathered; append $W$ to web_research_results.
    research_loop_count $\leftarrow$ research_loop_count $+ 1$.
    return $W$.
function SummarizeSources(state, $C$)
    Let $E \leftarrow$ running_summary (possibly empty).
    Let $W \leftarrow$ last element of web_research_results.
    if $E \neq$ "" then                                                      ▷ Update existing
        Build human message with $\langle$Existing Summary$\rangle E \langle/\rangle$ and $\langle$New Context$\rangle W \langle/\rangle$.
        Instruct: "Update the Existing Summary with the New Context on: research_topic."
    else                                                                     ▷ Create fresh
        Build human message with $\langle$Context$\rangle W \langle/\rangle$.
        Instruct: "Create a Summary using the Context on: research_topic."
    Call LLM with temperature 0 to get $S^\star$.
    if $C$.strip_thinking_tokens $= 1$ then strip $\langle$think$\rangle$ blocks in $S^\star$.
    running_summary $\leftarrow S^\star$; return $S^\star$.
function ReflectOnSummary(state, $C$)
    Provide running_summary and research_topic to LLM with a reflection prompt.
    if $C$.use_tool_calling $= 1$ then
        Request tool payload $\{$follow_up_query, knowledge_gap$\}$.
    else
        Request JSON $\{$follow_up_query, knowledge_gap$\}$.
    Extract $q' \leftarrow$ follow_up_query; on failure, strip $\langle$think$\rangle$ then fallback to "Tell me more about
        research_topic".
    search_query $\leftarrow q'$; return $q'$.
function FinalizeSummary(state)
    Initialize seen$\leftarrow \emptyset$, unique_lines$\leftarrow$ [ ].
    for all bullet_block $\in$ sources_gathered do
        for all line $\in$ lines(bullet_block) do
            if line is nonempty $\wedge$ line $\notin$ seen then
                add line to unique_lines; seen$\leftarrow$seen$\cup\{$line$\}$.
    all_sources$\leftarrow$ join(unique_lines, "\n").
    running_summary $\leftarrow$ "## Summary" || running_summary || "\n\n ### Sources:\n" || all_sources.
```

```
    return running_summary.
Main flow
search_query ← GenerateQuery($state, C$).
loop
    _ ← WebResearch($state, C$).
    _ ← SummarizeSources($state, C$).
    search_query ← ReflectOnSummary($state, C$).
    if research_loop_count $>$ $C$.max_web_research_loops then break
final_summary ← FinalizeSummary($state$).
return final_summary.
```

Algorithm 2: Local Evidence-First DR Agent

```
Require: Inputs: research_topic $\in \mathbb{S}$; configuration $C = \{$ llm_provider$\in \{$lmstudio, ollama$\}$,
         search_api$\in$ $\{$tavily, perplexity, duckduckgo, searxng$\}$, fetch_full_page$\in$ $\{0, 1\}$,
         strip_thinking_tokens$\in$ $\{0, 1\}$, use_local_rag$\in$ $\{0, 1\}$, vector_store_paths: List,
         embedding_model$\in$ $\mathbb{S}$, local_results_count$\in$ $\mathbb{N}$, research_mode$\in$ $\{$single, DToR$\}$,
         max_web_research_loops$\in$ $\mathbb{N}\}$.
Ensure: final running_summary; in single mode it includes deduplicated sources (tagged by chan-
        nel), in DToR mode it omits sources.
State: running_summary ← ""; summary_history ← [ ]; search_query ← "";
complementary_search_query ← ""; local_research_results ← [ ]; local_rag_summary ←
""; web_research_results ← [ ]; complementary_web_research_results ← [ ];
local_sources_gathered ← [ ]; sources_gathered ← [ ]; complementary_sources_gathered ← [ ];
research_loop_count ← 1.
function GenerateQuery(state, $C$)
    Build system prompt from research_topic and current_date.
    Gate JSON mode by provider/model; attempt JSON parse of {query,rationale}.
    if JSON parse fails then
        if $C$.strip_thinking_tokens $= 1$ then strip ⟨think⟩
        Set $q$ ← first nonempty line of content, else $q$ ← research_topic.
    search_query ← $q$; return $q$.
function LocalRAGResearch(state, $C$)
    if $C$.use_local_rag $= 0$ then return {local_research_results : [ ], local_sources_gathered :
    [ ]}
    Expand $C$.vector_store_paths into discovered volumes; query Chroma stores with device-
    aware embeddings.
    Extract raw contents list $R$; format citations $S_{\mathrm{loc}}$.
    return {local_research_results : $R$, local_sources_gathered : [$S_{\mathrm{loc}}$]}.
function SummarizeLocalRAGResults(state, $C$)
    if local_research_results $= \emptyset$ then return {local_rag_summary : ""}
    Join local results into one string; prompt concise ($< 500$ chars) summary with fixed lead-in.
    return {local_rag_summary : $\ell$}.
function GenerateComplementaryQuery(state, $C$)
    if $C$.use_local_rag $= 0$ or local_research_results $= \emptyset$ then
        return {complementary_search_query : ""}            ▷ skip when no local signal
    Ask LLM (JSON) using local_rag_summary context; recursively search any JSON key
    containing "query".
    if extracted query empty $\vee$ identical to search_query then
        apply backup prompt(s) and safe template fallbacks.
    return {complementary_search_query : $q_c$}.
function WebResearch(state, $C$)
    Call configured backend with $q$ = search_query.
    Build $W$ ← deduplicate_and_format_sources(·, max_tokens_per_source =
    2500, $C$.fetch_full_page).
    Append formatted bullets to sources_gathered and $W$ to web_research_results.
    research_loop_count ← research_loop_count + 1; return $W$.
function ComplementaryWebResearch(state, $C$)
    if complementary_search_query = "" then
        return {complementary_sources_gathered : [ ], complementary_web_research_results :
        [ ]}
```

```
    Query backend with $q_c$; build $W_c$ as above (budget 2500 tokens per source).
    return {complementary_sources_gathered : $[S_c]$, complementary_web_research_results : $[W_c]$}.
function SummarizeSources(state, $C$)
    Let $E \leftarrow$ running_summary; assemble context string $X$ by concatenating
        (i) last of local_research_results if any as ⟨Local⟩ . . . ⟨/⟩,
        (ii) last of web_research_results as ⟨Main⟩ . . . ⟨/⟩,
        (iii) last of complementary_web_research_results as ⟨Complementary⟩ . . . ⟨/⟩.
    if $E \neq$ ”” then prompt to update $E$ with $X$;
    elseprompt to create summary from $X$.
    Get $S^\star$ (temperature 0.3); if $C$.strip_thinking_tokens $= 1$ then strip tokens.
    Append {iteration = research_loop_count, summary = $S^\star$, query = search_query, complementary_query = complementary_search_query} to summary_history.
    running_summary $\leftarrow S^\star$; return $S^\star$.
function ReflectOnSummary(state, $C$)
    Request JSON with keys {follow_up_query, knowledge_gap}; use model-specific JSON gating.
    if parse fails then
        if $C$.strip_thinking_tokens $= 1$ then strip ⟨think⟩
        fallback to a safe heuristic query.
    search_query $\leftarrow q'$; return $q'$.
function RouteResearch(state, $C$)
    if research_loop_count $\leq C$.max_web_research_loops then return local_rag_research
    else return finalize_summary
function FinalizeSummary(state, $C$)
    Build history text from summary_history if available, else use running_summary.
    Ask LLM (temperature 0.3) for final mini-report.
    if $C$.research_mode = DToR then return final running_summary ▷ sources handled at DToR synthesis
    else
        Deduplicate lines from sources_gathered (main), complementary_sources_gathered (tag as complementary), and local_sources_gathered (tag as local); join as ### Sources: block.
        Append sources block to final summary and return .
Main flow
INIT_SESSION; search_query $\leftarrow$ GenerateQuery$(state, C)$.
LocalRAGResearch(state,C) → SummarizeLocalRAGResults(state,C) → GenerateComplementaryQuery(state,C).
WebResearch(state,C) → ComplementaryWebResearch(state,C) → SummarizeSources(state,C).
search_query $\leftarrow$ ReflectOnSummary$(state, C)$.
if RouteResearch(state,C) = local_rag_research then goto LocalRAGResearch else FinalizeSummary(state,C).
```

Algorithm 3: DToR Orchestrator Over Research Nodes (Breadth–Depth Controller)

```
Require: Inputs: research_topic $\in \mathbb{S}$; configuration $C = \{$ llm_provider, search_api, fetch_full_page, strip_thinking_tokens, use_local_rag, vector_store_paths, embedding_model, max_branches$\in \mathbb{N}$, max_branch_depth$\in \mathbb{N}$, nodes_per_branch$\in \mathbb{N}\}$.
Ensure: final_summary and (optionally) all_sources.
State (ToTState):
    branches: Dict[branch_id → ResearchBranch], where each
        ResearchBranch = {perspective, depth, remaining_budget, research_nodes: List<SummaryState>, branch_summary, is_complete};
    active_branch_id; knowledge_gaps; all_sources; is_complete $\in \{0, 1\}$.
function DiversifyInitialQuery(state, $C$)                              ▷ Breadth seeding
    Generate up to $C$.max_branches orthogonal perspectives (titles, descriptions, seed queries).
    For each perspective $i$: create a branch with one RN (SummaryState) seeded by its query;
    set depth $\leftarrow 0$, remaining_budget $\leftarrow C$.nodes_per_branch.
```

```
    Choose initial active_branch_id.
function SELECTNEXTBRANCH(state)
    if active branch has actionable work then return active_branch_id
    elsereturn first incomplete branch id
function RUNRESEARCHNODE(state, $C$)                ▷ Depth step = RN from Alg. 2
    In the active branch, pick the next RN with processing_status = "pending".
    Invoke Algorithm 2 on that RN (local evidence $\rightarrow$ web $\rightarrow$ summarize $\rightarrow$ reflect).
    Update the RN with its running_summary and mark completed; clear per-node sources in DToR mode.
function DECISIONANALYST(state, $C$)                ▷ EXPAND or PRUNE only
    Read the latest completed RN in the active branch and its context (depth, remaining_budget).
    Return decision $\in \{\text{EXPAND}, \text{PRUNE}\}$ and (if EXPAND) a small list of knowledge_gaps with web queries.
    if depth $\geq C$.max_branch_depth then
        force decision $\leftarrow$ PRUNE; knowledge_gaps $\leftarrow [\,]$
    Mark the analyzed RN as processing_status = "analyzed".
function GENERATEFOLLOWUPQUERIES(state, $C$)        ▷ Depth expansion
    Let $k \leftarrow \min(|\text{knowledge\_gaps}|,\ \text{remaining\_budget})$.
    For the first $k$ gaps: create new RN nodes with refined topics/queries (avoid compounded prefixes).
    remaining_budget $\leftarrow$ remaining_budget $- k$; depth $\leftarrow$ depth $+ 1$.
function SYNTHESIZEBRANCH(state)
    Compress all RN running_summary in the branch into branch_summary;
    deduplicate and append sources into all_sources; persist branch report.
function SYNTHESIZEFINAL(state)
    Integrate all branch_summary across branches into final_summary; mark is_complete $\leftarrow 1$.
function ROUTENEXTACTION(state, $C$)                ▷ Breadth–depth control
    if not in DToR mode or state.is_complete then return END
    if no branches then return DIVERSIFYINITIALQUERY
    Let $b \leftarrow$ active branch.
    if knowledge_gaps $\neq \emptyset$ and $b$.remaining_budget $> 0$ then return GENERATEFOLLOWUPQUERIES
    if exists pending RN in $b$ then return RUNRESEARCHNODE
    if exists completed but not analyzed RN in $b$ then return DECISIONANALYST
    if $b$.is_complete $= 1$ and $b$.branch_summary empty then return SYNTHESIZEBRANCH
    if all branches complete and summarized then return SYNTHESIZEFINAL
    return SELECTNEXTBRANCH
Main flow (controller loop)
DIVERSIFYINITIALQUERY(state,$C$); active_branch_id $\leftarrow$ SELECTNEXTBRANCH($state$).
while not state.is_complete do
    $a \leftarrow$ ROUTENEXTACTION($state, C$).
    if $a =$ RUNRESEARCHNODE then RUNRESEARCHNODE(state,C)
    if $a =$ DECISIONANALYST then
        (decision, knowledge_gaps) $\leftarrow$ DECISIONANALYST($state, C$).
        if decision = EXPAND and $b$.depth $< C$.max_branch_depth then GENERATEFOLLOWUPQUERIES(state,C)
        else                                        ▷ PRUNE $\equiv$ stop expanding this branch
            $b$.is_complete $\leftarrow 1$
    if $a =$ GENERATEFOLLOWUPQUERIES then GENERATEFOLLOWUPQUERIES(state,C)
    if $a =$ SYNTHESIZEBRANCH then SYNTHESIZEBRANCH(state)
    if $a =$ SYNTHESIZEFINAL then SYNTHESIZEFINAL(state); break
    if $a =$ SELECTNEXTBRANCH then active_branch_id $\leftarrow$ SELECTNEXTBRANCH($state$)
return final_summary (and all_sources).
```

D Supplemental Results: Experimental Setup

D.1 Nanomaterials/Devices Topic Suite

PFAS FET Sensor Probe

Which two-dimensional (2D) nanomaterials or molecular probes-including novel, underexplored candidates-offer the greatest potential to achieve highly sensitive and selective detection of per- and polyfluoroalkyl substances (PFAS), or more broadly chemical and biological analytes, when integrated into a field-effect transistor (FET)-based sensor platform?

Battery Binder

Which advanced binder technologies are being developed to improve the performance and longevity of lithium-ion batteries?

Staphylococcus aureus FET Biosensor

How can emerging probe chemistries that are commercially available, cost-effective, and exhibit minimal batch-to-batch variability be used to develop a novel FET-based biosensor for detecting Staphylococcus aureus?

Resource Recovery Electrocatalysts

Which nanostructured electrocatalyst materials demonstrate the highest selectivity and efficiency for electrochemical detection and recovery of critical resources (e.g., Li^+, PO_4^{3-}, NH_4^+) from complex wastewater matrices, and what key performance metrics-such as selectivity, sensitivity, recovery rate, energy consumption, and operational stability-distinguish them?

OER Stability

Identify the top nanostructured electrocatalyst materials for driving the oxygen evolution reaction (OER) in complex wastewater matrices-such as high-chloride, high-organic-load, or multi-ion streams-, for each the material class/composition, comprehensively consider key performance metrics (overpotential at 10 mA cm^{-2}, Faradaic efficiency, stability), and the surface-engineering strategies that confer corrosion resistance and sustained activity in real effluent conditions.

Photothermal Wastewater Evaporation

What photothermal materials and system designs are most effective for solar-driven water evaporation in complex wastewater matrices-such as high-organic-load or multi-ion streams-and how do they compare in terms of material composition, solar-to-vapor conversion efficiency under one-sun illumination, evaporation rate, fouling resistance, and integrated resource-recovery functionalities?

Nutrient Sensor Interference

Which sensor probe materials/chemicals and designs offer the best performance in minimizing interference factors-such as competing ions, dissolved organic matter, pH fluctuations, and temperature variations-for accurate and selective detection of nutrients (e.g., nitrate, phosphate, ammonium) in complex water matrices, and what mitigation strategies do they employ?

CO2 Sensor Probe

Which two-dimensional materials-such as graphene derivatives, transition metal dichalcogenides, or MXenes-offer the highest CO_2 sensing performance in complex gas or aqueous environments, and how do they compare in terms of detection limit (ppm), selectivity against common interferents (e.g., O_2, H_2O), response/recovery time, and long-term stability, including any functionalization or structural modifications that enhance these metrics? Think of novel candidates.

Printed FET Variability

Which printed-electronics fabrication parameters and post-processing strategies-such as ink viscosity, printing speed and resolution, substrate surface energy, annealing temperature profiles, and in-line calibration methods-have been shown to minimize device-to-device variability in FET sensor arrays, and what specific process windows achieve low variation in threshold voltage and field-effect mobility?

2D Synaptic Transistors

Which two-dimensional material platforms (e.g., MoS_2, WSe_2, black phosphorus, h-BN), device architectures (e.g., floating-gate, ionic-gated, dual-gate), and fabrication protocols (e.g., channel thickness control, dielectric engineering, contact metallurgy) have been shown to optimize synaptic transistor performance-specifically in terms of energy per event, weight-update linearity, retention time, and cycling endurance-for neuromorphic sensing applications? Think of most practical and promising candidates.

Microplastics 2D Sensors

Which two-dimensional material platforms (e.g., graphene derivatives, transition metal dichalcogenides, MXenes), molecular recognition elements (e.g., molecularly imprinted polymers, aptamers, peptide receptors), and device integration strategies (e.g., FET, electrochemical impedance, photonic transduction) have demonstrated the highest sensitivity, selectivity against organic matter and ionic interferents, and rapid response times for detecting micro- and nanoplastic particles in complex water matrices? Think of novel candidates.

Antibiotics 2D Sensors

Which commercially available, cost-effective probe chemistries (e.g., thiolated DNA aptamers, antibody mimetics, conductive MIPs) combined with two-dimensional nanomaterial transducer platforms (e.g., WS_2 FETs, graphene field-effect sensors, nanotube-extended gate FETs) deliver the lowest detection limits, minimal batch variability, and robust performance for sensing trace levels of pharmaceutical antibiotics in diverse aqueous waterbody.

Li–Na Selective Membranes

What material or membrane exhibits the highest selectivity for Li^+ over Na^+ in aqueous systems, given their nearly identical hydrated ionic radii and solvation environments? Beyond crown-ether-functionalized polymers, what novel or unexpected materials-such as bioinspired ultrahigh-selectivity membranes or covalent organic framework nanochannels-might provide breakthrough Li^+ discrimination over Na^+? Find novel and promising candidates.

PFAS Degradation

Which novel electrode materials can achieve efficient PFAS degradation under ambient aqueous electrochemical conditions, delivering both high mineralization and defluorination rates? What intrinsic properties-such as PFAS adsorption affinity, reactive oxygen species generation capacity, and C-F bond activation energy-should be optimized to guide their discovery?

Li–Co–Ni Separation

Which membrane materials can effectively separate Li^+, Co^{2+}, and Ni^{2+} ions from aqueous solutions by leveraging selective transport properties-such as tailored pore sizes, specific surface functionalizations, and charge affinities-and what membrane design principles optimize both selectivity and permeability? Think of potential novel, effective, and practical candidates.

LC-TEM MoS2 FET Sensing

How can in situ liquid cell TEM be employed to directly visualize the real-time adsorption and structural changes of 2D MoS_2 nanosheets used in aqueous FET sensors during analyte binding? What fluid cell configurations and electron dose parameters are necessary to preserve native water-material interfaces while capturing high-resolution sensing events without beam-induced artifacts?

CO2RR Catalyst Leaders

Given the current landscape of CO_2 electroreduction, which state-of-the-art catalyst platforms-such as oxide-derived copper, single-atom catalysts on nitrogen-doped carbon supports, or metal-organic framework-derived materials-demonstrate the highest activity and selectivity? Moreover, what emerging catalytic systems or novel heterostructures beyond these examples could feasibly outperform today's leading electrocatalysts in terms of faradaic efficiency and stability?

Unconventional PV Platforms

Perovskite-silicon tandem cells, organic photovoltaics, and quantum-dot solar cells currently represent the forefront of next-generation solar technologies. Beyond these established platforms, which unconventional material classes or innovative device architectures-such as chalcogenide perovskites, 2D semiconductor heterostructures, or ferroelectric photovoltaic systems-offer the most unexpected promise for achieving breakthroughs in efficiency, stability, and scalability?

Ambient-Pressure Diamond Growth

How can ambient-pressure diamond synthesis using a Ga-Ni-Fe-Si liquid-metal alloy at 1 atm and ~1025 °C be adapted to produce larger-area diamond films or oriented single crystals, and what key mechanistic steps govern nucleation and growth kinetics under these mild-pressure conditions?

Diamane Property Tuning

What chemical functionalizations or reaction pathways are most promising for tuning the electronic bandgap and mechanical stability of diamane-like 2D diamond films formed via sp^2-to-sp^3 conversion, and how might those modifications influence their integration into nanoelectronic devices?

Anion-Receptor Ether Electrolytes

Search for novel fluorinated ether-based electrolyte candidates that incorporate covalently bound anion-receptor motifs (e.g. boron- or phosphorus-centered groups) to deliver oxidative stability beyond 5.6 V, Li^+ transference numbers above 0.8, and minimal interfacial impedance on lithium-metal anodes.

D.2 Details of DR Agent Experimental Settings

Open-source LLMs and local RAG corpus We implement the framework in LangGraph and deploy three low-deployment-cost, open-source LLMs—DeepSeek-R1-14B, DeepSeek-R1-70B, and QwQ-32B—served locally via Ollama. All agent prompts and tool I/O use a JSON/tool-calling discipline for determinism and easy parsing, and the same research node (RN) described in Algorithm 2 is the unit of work in the controller (Algorithm 3 DToR). To provision representative, application-driven knowledge for nanomaterials and devices, we build a standardized Chroma vector store spanning 1,136,297 full-text papers organized around four pillars: sensors, batteries, catalysis, and semiconductors. The corpus is queried first under a local-first RAG policy; complementary web retrieval is invoked only when Algorithm 2 (local evidence-first)'s reflection detects unresolved knowledge gaps. Consistent with our emphasis on report-level usefulness rather than information retrieval (IR) benchmarking, the local-first RAG layer is treated as a routine, swappable component; DToR only triggers complementary web evidence upon explicit knowledge gaps and resolves cross-source conflicts by evidence majority $\rightarrow$ recency $\rightarrow$ domain authority. To make domain coverage portable and auditable without exposing any local texts or indices, we provide exact domain-scoping query templates corresponding to these four pillars (sensors, energy/environment covering batteries, semiconductors, and catalysis) so others can target the same scope with their own holdings while keeping the orchestration identical.

Corpus construction Raw full texts are normalized with a cleaning pipeline designed for scientific documents at scale. The cleaner (i) preserves section structure and headers, (ii) retains LaTeX/MathJax equations and common math environments via stable markers, (iii) standardizes tables into a compact textual schema, (iv) normalizes URLs/DOIs and citation markers instead of stripping them, and (v) removes XML/HTML artifacts and noisy back matter conservatively. For long inputs, we apply sliding-window tokenization (default window 8192 tokens, stride 4096) or a section-aware variant that respects detected boundaries, yielding JSONL suitable for downstream training or indexing. As our focus is on the DToR orchestration, we adhere to widely used, off-the-shelf preparation and vector-indexing practices; the corpus layer functions as a routine, swappable component rather than a methodological contribution. To enable scope-equivalent replication across different holdings, we release the exact domain-scoping query templates for sensors, energy/environment (including batteries), semiconductors, and catalysis; users may instantiate a local store with their customary institutional pipelines and preferred indexers.

Indexing and retrieval Cleaned texts are split with a recursive character splitter (default chunk size 2500, overlap 500) and embedded using BAAI/bge-m3. Each chunk carries rich metadata (paper id, DOI, chunk id/index, source path, shard id) to support provenance and de-duplication. Embeddings are persisted in Chroma, enabling fast k-NN retrieval and incremental persistence during large ingestions. For robustness across machines, the store is partitioned into 500 volumes (shards). This organization serves two purposes: (1) controllable breadth—DToR can expand to additional volumes when the analyst decides to broaden coverage—and (2) bounded depth/latency—branches can be pruned once depth or evidence sufficiency is reached, without touching unrelated volumes. To mitigate licensing and privacy risk, our local corpus is restricted to institutionally licensed or open sources, and we release only the domain-scoping query templates—not texts or indices—so others can replicate scope without redistributing copyrighted content or PII.

Policy integration In each research node, local retrieval is attempted first; summaries are synthesized from local evidence. Only when reflection surfaces concrete gaps does the controller expand (selecting additional volumes and spawning new RNs) or prune (stopping further growth of a branch). This strategy couples breadth (multi-volume exploration across pillars) with depth (iterative drilling within a volume) while keeping costs predictable on commodity hardware and avoiding external dependencies during the main research loop.

Agent Configuration By default, we run all agents in a local-first configuration with cost-aware web fallback. Concretely, both single-agent DR and DToR branches use a research depth of 3 iterations, each RN queries the local Chroma first with top-k = 5, and only then invokes web search—by default DuckDuckGo—to avoid API fees. In DToR mode, breadth–depth is bounded with max branches = 3, max branch depth = 3, nodes per branch = 100, and a global cap of 500 nodes per experiment, ensuring predictable latency and compute. Models are served locally via Ollama; RAG is enabled by default with BAAI/bge-m3 embeddings, and we preserve document fidelity via full-page fetch and safe output hygiene. All parameters remain user-tunable through the configuration/RunnableConfig to trade off coverage, latency, and cost—for example, increasing depth or top-k for harder queries, or switching web backends when budgets permit.

D.3 Rubric Template for LLM-as-Judge Evaluation Experiment

LLM-as-Judge Rubric Prompt (applied to *DR Report*)

{Anonymous DR Report content}
Now please analyze and evaluate the Report above on five equally weighted dimensions:

1. Relevance to the stated objective,
2. Depth and rigor of validation and benchmarking protocols,
3. Clarity and reproducibility of methodology,
4. Practical applicability and implementation readiness,
5. Novelty and breadth of candidate coverage.

For each dimension, assign a score from 0–10 and supply a one-sentence justification. This rubric places equal emphasis on procedural rigor and clear implementation steps alongside overall alignment, ensuring that reports with well-defined benchmarks and actionable guidance receive full credit.

Example raw response (DToR_qwq32B_local500; Topic: PFAS FET sensor probe) — Claude 4 Opus (thinking)

```
Based on my analysis of the research report on 2D nanomaterials and molecular
↪  probes for FET-based PFAS detection, here are the evaluation scores across
↪  the five equally weighted dimensions:

## **Dimension 1: Relevance to the Stated Objective**
**Score: 9/10**

The report directly addresses the core question about identifying 2D
↪  nanomaterials and molecular probes with the greatest potential for highly
↪  sensitive and selective PFAS detection in FET-based sensors. The
↪  comprehensive analysis covers both established and novel candidates
↪  including MXenes, graphene, TMDCs, aptamers, and molecularly imprinted
↪  polymers, with specific focus on their integration into FET platforms for
↪  PFAS detection.

## **Dimension 2: Depth and Rigor of Validation and Benchmarking Protocols**
**Score: 6/10**

While the report provides specific performance metrics such as sensitivity
↪  ranges (100 fM to 10 pM) and specificity percentages (99% for PFOS), it
↪  lacks detailed descriptions of the experimental protocols, standardized
↪  testing conditions, and validation methodologies used to obtain these
↪  benchmarks. The report acknowledges this gap by calling for "standardized
↪  protocols for probe-material interface characterization" in its future
↪  research recommendations.

## **Dimension 3: Clarity and Reproducibility of Methodology**
**Score: 7/10**

The report presents a well-structured synthesis approach with clear
↪  categorization of materials, probes, and integration strategies, but
↪  provides limited detail on the specific research methodologies used to
↪  generate the findings from the five research branches (A-E). The
↪  contradiction analysis and trade-off evaluations demonstrate systematic
↪  thinking, though the underlying experimental procedures for material
↪  characterization and performance testing are not explicitly described.

## **Dimension 4: Practical Applicability and Implementation Readiness**
**Score: 8/10**
```

```
The report excels in addressing real-world implementation challenges,
↪  specifically highlighting manufacturing scalability through roll-to-roll
↪  processing, environmental stability concerns for MXenes, and the need for
↪  protective coatings. The comprehensive candidate inventory with performance
↪  metrics and the identification of hybrid systems like MXene/graphene
↪  heterostructures provide actionable pathways for practical deployment.

## **Dimension 5: Novelty and Breadth of Candidate Coverage**
**Score: 8/10**

The report demonstrates excellent breadth by covering both conventional
↪  materials (graphene, TMDCs) and emerging candidates (MXenes, fluorinated
↪  phospholipid bilayers), while exploring novel integration approaches and
↪  hybrid systems. The cross-disciplinary innovations section introduces
↪  creative solutions like machine learning optimization and bioinspired
↪  coatings, though some candidates like fluorinated phospholipid bilayers are
↪  noted as requiring further validation.

## **Overall Assessment**

**Total Score: 38/50 (76%)**

The report provides a comprehensive and well-structured analysis that
↪  effectively addresses the stated objective with strong practical focus and
↪  broad candidate coverage, but would benefit from more detailed validation
↪  protocols and methodological transparency to enhance reproducibility and
↪  scientific rigor.
```

Example raw response (DToR_qwq32B_local500; Topic: PFAS FET sensor probe) — Gemini 2.5 Pro

```
Based on the provided research report, here is an analysis and evaluation across
↪  the five requested dimensions:

*   **Relevance to the stated objective**
    *   **Score**: 10/10
    *   **Justification**: The report directly and comprehensively answers the
    ↪  core research question by identifying and evaluating specific 2D
    ↪  materials and molecular probes for FET-based PFAS detection.

*   **Depth and rigor of validation and benchmarking protocols**
    *   **Score**: 7/10
    *   **Justification**: While the report synthesizes key quantitative
    ↪  performance metrics, it correctly identifies the lack of standardized
    ↪  testing protocols as a critical gap in the field.

*   **Clarity and reproducibility of methodology**
    *   **Score**: 6/10
    *   **Justification**: The report clearly outlines high-level integration
    ↪  strategies but lacks the detailed experimental parameters required for
    ↪  direct methodological reproduction.

*   **Practical applicability and implementation readiness**
    *   **Score**: 9/10
    *   **Justification**: The analysis heavily focuses on practical barriers to
    ↪  commercialization, such as stability and scalability, and proposes a
    ↪  clear roadmap for real-world implementation.

*   **Novelty and Breadth of Candidate Coverage**
```

```
    *   **Score**: 10/10
    *   **Justification**: The report covers a wide breadth of candidates,
    ↪  including novel materials like MXenes and underexplored probes like
    ↪  fluorinated phospholipid bilayers, while also considering innovative
    ↪  hybrid systems.
```

Example raw response (DToR_qwq32B_local500; Topic: PFAS FET sensor probe) — Grok 3 (thinking)

```
Below is the evaluation of the provided research report titled *"Integrated
↪  Research Report: 2D Nanomaterials and Molecular Probes for FET-Based PFAS
↪  Detection"* based on the five equally weighted dimensions specified in the
↪  rubric. Each dimension is scored from 0 to 10, with a one-sentence
↪  justification for each score.

---

### **Evaluation of the Research Report**

#### **1. Relevance to the Stated Objective**
- **Score**: 9
- **Justification**: The report directly addresses the detection of PFAS and
↪  other analytes using FET-based sensors with 2D nanomaterials and molecular
↪  probes, covering material selection, probe design, and integration
↪  strategies; however, it could have explored the specific challenges of
↪  detecting diverse PFAS compounds in greater depth.

#### **2. Depth and Rigor of Validation and Benchmarking Protocols**
- **Score**: 7
- **Justification**: The report provides specific performance metrics such as
↪  sensitivity (100 fM-10 pM) and specificity (99%), but it lacks detailed
↪  descriptions of validation methods or comparisons with existing
↪  technologies, limiting the rigor of its claims.

#### **3. Clarity and Reproducibility of Methodology**
- **Score**: 8
- **Justification**: The methodology is clearly outlined for integration
↪  strategies like layer-by-layer assembly and roll-to-roll manufacturing, but
↪  it falls short on detailed, reproducible steps for material
↪  functionalization or probe design.

#### **4. Practical Applicability and Implementation Readiness**
- **Score**: 7
- **Justification**: The report discusses scalability through techniques like
↪  roll-to-roll manufacturing and identifies challenges, but it lacks
↪  actionable details such as cost estimates or specific implementation steps,
↪  keeping it somewhat theoretical.

#### **5. Novelty and Breadth of Candidate Coverage**
- **Score**: 10
- **Justification**: The report stands out for its comprehensive coverage of 2D
↪  materials (e.g., MXenes, graphene, TMDCs, BN) and molecular probes (e.g.,
↪  aptamers, MIPs, fluorinated phospholipid bilayers), including innovative
↪  concepts like hybrid systems and machine learning optimization.

---

### **Summary of Scores**
- **Relevance to the Stated Objective**: 9
- **Depth and Rigor of Validation and Benchmarking Protocols**: 7
- **Clarity and Reproducibility of Methodology**: 8
- **Practical Applicability and Implementation Readiness**: 7
```

```
- **Novelty and Breadth of Candidate Coverage**: 10

---

### **Overall Assessment**
The report is highly relevant to the objective of identifying 2D nanomaterials
↪  and molecular probes for FET-based PFAS detection and excels in its
↪  innovative and broad coverage of candidates. However, it would benefit from
↪  stronger validation protocols, including detailed benchmarking, and more
↪  specific guidance on practical implementation to enhance its real-world
↪  applicability. These improvements would bolster its procedural rigor and
↪  readiness while preserving its strengths in relevance and novelty.
```

Example raw response (DToR_qwq32B_local500; Topic: PFAS FET sensor probe) — ChatGPT o3

```
Overall, the report captures the current consensus that MXene- and
↪  aptamer-functionalised FETs can drive PFAS detection into the femto- to
↪  picomolar regime, mirroring recent experimental demonstrations of ultra-low
↪  limits of detection and hybrid material strategies.
↪  ([pmc.ncbi.nlm.nih.gov][1], [mdpi.com][2])  However, it gives only a
↪  qualitative treatment of validation and field-deployment protocols that
↪  agencies such as the U.S. EPA and DoD now require, leaving practical
↪  benchmarking details thin. ([epa.gov][3], [denix.osd.mil][4])  The table
↪  below assigns balanced scores across the five requested dimensions.

## Evaluation Table

| Dimension                                                | Score (0-10) |
↪  One-sentence justification
↪  |
| -------------------------------------------------------- | ------------ |
↪  ----------------------------------------------------------------------------⌋
↪  ----------------------------------------------------------------------------⌋
↪  ----------------------------------------------------------------------------⌋
↪  -----------------------------------------------------------------------------
↪  |
| **1. Relevance to Objective**                            | **9**        | The
↪  discussion centres on MXenes, graphene, TMDCs and probe chemistries
↪  (aptamers, MIPs) that literature identifies as the most sensitive FET
↪  transducers for PFAS and analogous trace contaminants, directly matching the
↪  stated goal. ([pmc.ncbi.nlm.nih.gov][1], [mdpi.com][5],
↪  [pmc.ncbi.nlm.nih.gov][6]) |
| **2. Depth & Rigor of Validation/Benchmarking**          | **6**        |
↪  While impressive detection limits are quoted, the report lacks specific
↪  inter-laboratory comparisons, LOD calculation methods, or reference to
↪  emerging EPA/DoD validation workflows, so its rigor is only moderate.
↪  ([epa.gov][3], [denix.osd.mil][4])
↪  |
| **3. Clarity & Reproducibility of Methodology**          | **7**        | A
↪  clear LbL/R2R/CVD integration scheme is laid out, but omission of key
↪  experimental parameters (e.g., gate bias, buffer composition, chip layout)
↪  hampers full reproducibility expected in contemporary FET biosensor reports.
↪  ([exploration-journals.com][7], [pubs.rsc.org][8])
↪  |
| **4. Practical Applicability & Implementation Readiness** | **6**        | The
↪  narrative recognises roll-to-roll graphene films and protective coatings to
↪  curb MXene oxidation, yet stops short of cost modelling, encapsulation, or
↪  shelf-life testing that determine commercial readiness.
↪  ([link.springer.com][9], [pubs.rsc.org][10], [sciencedirect.com][11])
↪  |
```

```
| **5. Novelty & Breadth of Candidate Coverage**            | **8**        |
↪  Inclusion of under-explored fluorinated bilayers, machine-learning-optimised
↪  aptamers, and MXene/graphene heterostructures shows commendable breadth
↪  beyond mainstream materials and highlights emerging cross-disciplinary
↪  avenues. ([nature.com][12], [sciencedirect.com][13], [mdpi.com][2])
↪  |

### Key Take-aways

* **Strengths:** High topical relevance, comprehensive material/probe coverage,
↪  and acknowledgment of scalability challenges.
* **Gaps to Address:** Explicit validation protocols, detailed experimental
↪  conditions, and economic/packaging analyses would improve rigor and
↪  real-world impact.

By closing these gaps, the report could rise to a top-tier reference for
↪  PFAS-oriented FET sensor development.

[1]: https://pmc.ncbi.nlm.nih.gov/articles/PMC12109899/?utm_source=chatgpt.com
↪  "Aptamer-Based Planar Electric Double-Layer Field-Effect Transistor"
[2]: https://www.mdpi.com/2079-6374/14/10/497?utm_source=chatgpt.com "Recent
↪  Advancements in MXene-Based Biosensors for Health and ..."
[3]: https://www.epa.gov/water-research/pfas-analytical-methods-development-and⌋
↪  -sampling-research?utm_source=chatgpt.com "PFAS Analytical Methods
↪  Development and Sampling Research"
[4]: https://www.denix.osd.mil/edqw/denix-files/sites/43/2023/02/Module-6-Data-⌋
↪  Validation-Guidelines-1633-PFAS-Final-1.pdf?utm_source=chatgpt.com "[PDF]
↪  Module 6 Data Validation Guidelines - 1633 PFAS Final - DOD DENIX"
[5]: https://www.mdpi.com/2227-9040/12/7/116?utm_source=chatgpt.com "First
↪  Direct Gravimetric Detection of Perfluorooctane Sulfonic Acid ..."
[6]: https://pmc.ncbi.nlm.nih.gov/articles/PMC7944395/?utm_source=chatgpt.com
↪  "MXene-Graphene Field-Effect Transistor Sensing of Influenza Virus ..."
[7]: https://www.exploration-journals.com/article/aptamer-functionalized-field-⌋
↪  effect-transistor-biosensors-for-disease-diagnosis-and-environmental-monito⌋
↪  ring/?utm_source=chatgpt.com "Aptamer-functionalized field-effect transistor
↪  biosensors for disease ..."
[8]: https://pubs.rsc.org/en/content/articlehtml/2022/ma/d1ma00929j?utm_source=⌋
↪  chatgpt.com "Microfluidic sensors based on two-dimensional materials for
↪  ..."
[9]: https://link.springer.com/article/10.1007/s40820-024-01534-x?utm_source=ch⌋
↪  atgpt.com "Flexible Graphene Field-Effect Transistors and Their Application
↪  in ..."
[10]: https://pubs.rsc.org/en/content/articlehtml/2024/tc/d4tc00884g?utm_source⌋
↪  =chatgpt.com "Shielding 2D MXenes against oxidative degradation - RSC
↪  Publishing"
[11]: https://www.sciencedirect.com/science/article/abs/pii/S1385894724055864?u⌋
↪  tm_source=chatgpt.com "A review of how to improve Ti3C2Tx MXene stability -
↪  ScienceDirect"
[12]: https://www.nature.com/articles/s41467-021-22555-9?utm_source=chatgpt.com
↪  "Machine learning guided aptamer refinement and discovery - Nature"
[13]: https://www.sciencedirect.com/science/article/abs/pii/S0360319921047455?u⌋
↪  tm_source=chatgpt.com "Heterostructured two dimensional materials of MXene
↪  and graphene ..."
```

Example raw response (DToR_qwq32B_local500; Topic: PFAS FET sensor probe) — ChatGPT o4 mini high

```
Here are the scores and one-sentence justifications for each dimension:

## 1. Relevance to the Stated Objective

**Score: 9/10**
```

```
The report aligns strongly with the goal of identifying 2D materials and
↪  molecular probes for PFAS detection in FET sensors, covering both
↪  sensitivity and selectivity considerations, with only minor omissions around
↪  alternative non-FET platforms.

## 2. Depth and Rigor of Validation and Benchmarking Protocols

**Score: 6/10**
While key performance metrics (e.g., detection limits, specificity) are cited,
↪  the report lacks detailed descriptions of experimental controls, statistical
↪  analyses, and comparative benchmarks across different sensing platforms.

## 3. Clarity and Reproducibility of Methodology

**Score: 7/10**
Integration strategies and trade-offs are clearly delineated, but the absence of
↪  step-by-step protocols (e.g., exact functionalization procedures, probe
↪  immobilization conditions) limits precise reproduction.

## 4. Practical Applicability and Implementation Readiness

**Score: 8/10**
The discussion of scalability (R2R, LbL, CVD) and durability solutions offers
↪  actionable guidance for commercialization, though cost-analysis and
↪  regulatory considerations could be more fully addressed.

## 5. Novelty and Breadth of Candidate Coverage

**Score: 8/10**
The inventory spans established (graphene, TMDCs) and emerging (fluorinated
↪  bilayers, MXENE hybrids) candidates, but could further explore under-studied
↪  2D systems (e.g., phosphorene, borophene) and alternative probe chemistries.
```

Table 1: Example raw response (DToR_qwq32B_local500; Topic: PFAS FET sensor probe; Trial #1) — cross-model score matrix (0–10).

Model	**Relevance**	**Depth**	**Clarity**	**Applicability**	**Novelty**	**Overall (mean)**
Claude 4 Opus (thinking)	**9**	**6**	**7**	**8**	**8**	**7.6**
Gemini 2.5 Pro	**10**	**7**	**6**	**9**	**10**	**8.4**
Grok 3 (thinking)	**9**	**7**	**8**	**7**	**10**	**8.2**
ChatGPT o3	**9**	**6**	**7**	**6**	**8**	**7.2**
ChatGPT o4 mini high	**9**	**6**	**7**	**8**	**8**	**7.6**
Average across models	**9.2**	**6.4**	**7.0**	**7.6**	**8.8**	**7.8**

A/B Dueling Prompt (*DR Report A* vs. *DR Report B*)

Report A: {Anonymous DR Report content}
Report B: {Anonymous DR Report content}
As an experienced materials scientist who bridges the gap between research and pilot lines, please review these two reports. I need your professional judgment: which one offers the clearer high-level strategic roadmap for our team's R&D over the next six months? We are especially interested in the report that presents a well-structured narrative, systematically identifies and prioritizes knowledge gaps, and synthesizes actionable next steps—beyond performance metrics. Please give me your recommendation in a brief, plain-language note—straight to the point—ending with a clear verdict on which report to back and the single most decisive reason why.

A/B Duel — Metadata

```
agent_a: DToR_qwq32B_local500_RD3_RAG5
agent_b: ChatGPT_o3_DR
trial: trial_1
```

Raw dueling response (DToR_qwq32B_local500; Topic: PFAS FET sensor probe)— Claude 4 Opus (thinking)

```
**Professional Assessment: Report A vs Report B**

After reviewing both reports as a materials scientist with pilot-line
↪  experience, **Report A is the clear choice** for your R&D roadmap.

**Why Report A wins:**

Report A provides a structured, decision-ready framework that directly addresses
↪  your team's needs. It synthesizes findings across five research branches,
↪  explicitly identifies contradictions (like MXene sensitivity vs. stability),
↪  and most importantly, provides **resolution strategies** for each challenge.
↪  The report prioritizes MXene-aptamer systems while acknowledging their
↪  limitations and offering concrete solutions like hybrid coatings and R2R
↪  manufacturing approaches.

Report A also includes a practical "Candidate Inventory" table that ranks
↪  materials by performance metrics and a clear action plan focusing on three
↪  critical barriers: stability testing, standardized protocols, and
↪  cross-disciplinary collaboration.

**Why Report B falls short:**

Report B is comprehensive but reads more like a literature review than a
↪  strategic roadmap. While it contains valuable technical depth, it lacks
↪  prioritization and actionable synthesis. The extensive material descriptions
↪  don't clearly guide decision-making, and the conclusions are more
↪  observational than directive.

**The decisive factor:** Report A explicitly addresses the "how to move forward"
↪  question with ranked recommendations and specific next steps, while Report B
↪  provides extensive "what exists" information without clear strategic
↪  direction.

**Verdict: Back Report A** - it's the roadmap your R&D team can immediately act
↪  upon.
```

Raw dueling response (DToR_qwq32B_local500; Topic: PFAS FET sensor probe)— Gemini 2.5 Pro

```
To the R&D Lead,

After reviewing both documents, my recommendation is to use **Report A** as the
↪  strategic roadmap for your team's R&D efforts over the next six months.

While Report B is an excellent and exhaustive technical review, Report A is
↪  structured as an executive-level strategic plan. It moves beyond a simple
↪  survey of materials and performance metrics to actively synthesize the
↪  information into a coherent strategy. It identifies the most critical
↪  trade-offs (e.g., Sensitivity vs. Stability), frames them as solvable
↪  problems in its "Contradiction Analysis," and provides a clear, prioritized
↪  conclusion with actionable next steps for your team.

My verdict is to back **Report A**.
```

```
The single most decisive reason is: **Report A is a decision-making tool, not
↪  just a literature review; its structure forces a strategic focus on
↪  resolving key conflicts, which is exactly what a short-term R&D plan
↪  requires.**
```

Raw dueling response (DToR_qwq32B_local500; Topic: PFAS FET sensor probe)— Grok 3 (thinking)

```
After reviewing both reports as a materials scientist with experience in scaling
↪  R&D from lab to pilot, Report A stands out for its concise, integrated
↪  structure that weaves together multi-branch insights into a cohesive
↪  narrative. It systematically flags key gaps (e.g., stability in humid
↪  environments, scalability via manufacturing methods) and prioritizes them
↪  through contradiction resolutions and a focused conclusion, while distilling
↪  actionable steps like stability testing, protocol standardization, and
↪  collaborative innovations to guide near-term R&D.

Report B, while thorough and data-rich, reads more like a broad literature
↪  review with less emphasis on gap prioritization and explicit roadmapping,
↪  making it better for background reference than strategic planning.

**Verdict: Back Report A.** The single most decisive reason is its synthesis of
↪  actionable next steps into a clear, prioritized R&D framework that directly
↪  addresses real-world deployment barriers beyond just metrics.
```

Raw dueling response — ChatGPT o4 mini high

```
After reviewing both documents, Report A delivers a concise, goal-oriented
↪  roadmap that explicitly prioritizes knowledge gaps (e.g. MXene stability,
↪  scalable integration, standardized testing) and pairs each with actionable
↪  next steps (hybrid coatings development, R2R pilot runs, cross-disciplinary
↪  protocol design). In contrast, Report B, while rich in technical depth,
↪  reads more like an encyclopedic survey than a six-month R&D playbook.

Verdict: Back Report A-because it alone translates material insights into a
↪  structured, gap-driven action plan suitable for immediate project planning.
```

Raw dueling response (DToR_qwq32B_local500; Topic: PFAS FET sensor probe)— ChatGPT o3

```
Both reports demonstrate deep technical insight, but they serve different
↪  purposes.

• Report A reads like an executive playbook. It starts with a concise problem
↪  statement, then walks us through material/probe choices, trade-offs,
↪  contradictions, and cross-disciplinary innovations in a clearly labeled,
↪  logical sequence. Crucially, it flags three ‘‘blocking’’ knowledge gaps (MXene
↪  stability, manufacturing scale-up, and standardized testing) and ties each
↪  gap to concrete, near-term actions (e.g., hybrid coatings, roll-to-roll
↪  trials, protocol drafting). The Candidate Inventory table distills options
↪  at a glance and can be plugged directly into a six-month work-plan.

• Report B is an encyclopedic review. It is rich in context and citations, but
↪  the narrative meanders, mixes PFAS with many other analytes, and buries
↪  action items inside long technical passages. The sheer length makes it
↪  harder to extract priorities or sequence tasks.

Verdict: Back Report A-because it delivers a structured, priority-driven roadmap
↪  we can act on immediately.
```

E First-Principles Simulation

E.1 DFT simulation settings

Static Simulation

To simulate the binding interactions of diverse combinations under periodic boundary conditions, we employed ab initio calculations employing parameters optimized for precision and computational efficiency. The plane-wave energy cutoff was set to 400 eV, and spin-polarized density functional theory DFT with the Perdew-Burke-Ernzerhof (PBE) functional was utilized to describe exchange-correlation interactions. The Brillouin zone integration employed a Gamma centered $1 \times 1 \times 1$ k-point grid and van der Waals corrections were incorporated using the DFT-D3 method to capture dispersion effects. Ionic relaxations were conducted with a convergence criterion of 0.02 eV/Å for the forces, and the electronic structure iterations were set to converge to an energy difference of 10^{-4} eV. Implicit solvent field is added for static simulations by setting the dielectric constant at 78.4 if applicable (namely excluding the ab initio molecular dynamics (AIMD) with explicit solvent molecules, and the CO_2 sensor probe binding scenarios). The modeling of the slabs is directly conducted by human expert based on anonymously reading the DR report. To ensure fair comparison, we asked the expert to create same number of candidates from each report (e.g. 4 from DR report A, and 4 from DR report B), and also pick one domain commonly chosen benchmark material/chemical for each of the 5 applications. To prevent interactions between periodic images, we also include $>$20 Å vacuum gap along the z-, ensuring a reliable comparison between adsorbed and separated configurations.

Binding Energy Convention.

Following a standard DFT convention (Phys. Rev. B, 2008, 77, 12) for adsorption on typical studied 2D material systems like graphene, we report the binding energy ΔE_{bind} of an adsorbate at a given coverage as

$$\Delta E_{\text{bind}} = (E_{\text{clean\_slab}} + E_{\text{adsorbate}}) - E_{\text{adsorbed}} \tag{1}$$

where E_{adsorbed} is the total energy of the fully relaxed adsorbate–slab complex (for certain applications, bottom atoms of bulk crystals are frozen and surface area are free to mimic realistic surface interactions), $E_{\text{clean\_slab}}$ is the energy of the corresponding pristine clean slab supercell, and E_{adsorbed} is the energy of the isolated adsorbate in the same supercell. All terms are computed with identical exchange–correlation functional, plane-wave cutoff, and k-point mesh. Unless stated specifically (like for battery binder normalized to be contact areal) We report ΔE_{bind} for only one adsorbate in the slab super cell for qualitative comparison. This definition mirrors the adsorption-energy convention widely used for molecules on typical substrates like graphene or MXene (we adopt the term "binding energy" throughout). By this convention, a more positive ΔE_{bind} indicates an exothermic, favorable adsorption (energy released upon binding).

Ab Initio Molecular Dynamics (AIMD) Simulation

To improve the physical realism of our calculations and especially considering solvent effect for surface events, we embedded explicit water molecules in the supercell according to conformations identified in preceding static simulations (Mol. Syst. Des. Eng. 2025, 10, 345–356; Sci. Adv. 2025, 11, eadr9038) using an end-state protocol. Specifically, the optimized adsorbed configuration corresponding to E_{adsorbed} was used as the initial snapshot for the solvated *adsorbed* state, with explicit water molecules added thereafter. For the *pulled* state, the adsorbate was translated by at least 15 Å away from the relaxed slab (e.g., graphene)—without additional geometry optimization—and then solvated under identical conditions. These two equilibrated states form the basis for directly comparing adsorption energetics in explicit solvent, consistent with end-state comparisons commonly applied in MD umbrella-sampling studies (J. Phys. Chem. C 2023, 127, 5920–5930; ChemPhysChem 2018, 19, 690–702), and in line with benchmark explicit-solvent AIMD treatments of metal/water interfacial solvation (J. Chem. Phys. 2020, 152, 144703) and first-principles determinations of interfacial hydration free energies for adsorbates on Pt (J. Chem. Phys. 2021, 154, 094107).

The water density was maintained at 1 g cm^{-3} (for battery binder, solvent composition is different as describe in corresponding section). Our goal was to compare the evolution of the system's potential (electronic) energy upon equilibration (Electrocatalysis 2017, 8, 577–586; Commun. Chem. 2022, 5, 76; J. Chem. Phys. 2020, 152, 144703; J. Chem. Phys. 2021, 155, 134703).

To ensure that the solvent adopted physically meaningful configurations, each system was first equilibrated for 2 ps at 300 K with a 0.5 fs time step. Temperature was regulated via a Langevin thermostat to suppress extraneous inertial effects while preserving canonical sampling. After equilibrium, we further performed an 3 ps production run, from which we extracted time-averaged electronic potential energies.

We therefore report an explicit-solvent end-state potential-energy gap, denoted $\Delta E_{\mathrm{pot}}^{\mathrm{ES}}$ (AIMD), defined as the difference between the time-averaged electronic potential energies of the solvated *adsorbed* and *pulled* basins in identical cells:

$$\Delta E_{\mathrm{pot}}^{\mathrm{ES}} \equiv \langle E_{\mathrm{pot}} \rangle_{\mathrm{pulled}} - \langle E_{\mathrm{pot}} \rangle_{\mathrm{ads}}. \tag{2}$$

Unless otherwise noted, all comparisons use the same supercell, water count, ion composition, thermostatting, and slab-electrostatics settings. Hence, the more positive $\Delta E_{\mathrm{pot}}^{\mathrm{ES}}$ (AIMD), the more energetically favorable the adsorbed state—i.e., a potential-energy analogue of the binding energy—whereas a negative $\Delta E_{\mathrm{pot}}^{\mathrm{ES}}$ indicates that explicit solvation stabilizes the pulled state and adsorption is disfavored.

Scope and limitations of $\Delta E_{\mathrm{pot}}^{\mathrm{ES}}$:

$\Delta E_{\mathrm{pot}}^{\mathrm{ES}}$ is not a rigorously converged adsorption free energy. By construction it (i) neglects configurational and translational entropy; (ii) folds solvent reorganization and water (or other solvent)–surface exchange energetics into the potential-energy baseline; and (iii) depends on finite-time sampling of each basin rather than a reversible pathway or an alchemical thermodynamic cycle. Consequently, its absolute magnitudes can reflect not only adsorbate–surface interactions but also large-scale differences in hydrogen-bond networks and long-range electrostatics associated with hundreds of explicit waters. Nonetheless, because all systems are treated consistently within the same AIMD framework, $\Delta E_{\mathrm{pot}}^{\mathrm{ES}}$ serves as a heuristic yet informative comparative descriptor for explicit-solvent binding trends across molecules and candidate surfaces, revealing solvent-mediated, surface-dependent effects that are not captured in vacuum or implicit-solvent calculations. Formal free-energy methods (umbrella sampling/PMF, metadynamics, and double-decoupling/TI with standard-state and slab-electrostatics corrections) are acknowledged as the rigorous routes to ΔG, as exemplified for interfacial adsorbates at water–Pt interfaces (J. Chem. Phys. 2021, 154, 094107).

Computational feasibility rationale:

Rigorous free-energy calculations become computationally prohibitive for our systems, which routinely exceed 500 atoms and include numerous transition-metal centers in addition to light elements; fully converged interfacial free-energy landscapes are beyond feasible AIMD timescales under such conditions. Extensive explicit-solvent AIMD benchmarks of metal/water interfaces have likewise highlighted the costliness of such simulations (J. Chem. Phys. 2020, 152, 144703). In this context, the explicit-solvent end-state potential-energy gap provides a practical dynamic compromise: it captures key solvent-mediated effects absent from static DFT binding energies while remaining tractable for large, chemically realistic slabs. Thus, although approximate, $\Delta E_{\mathrm{pot}}^{\mathrm{ES}}$ offers higher fidelity and greater interpretive value than purely static vacuum or implicit-solvent descriptions for the screening-level comparisons emphasized in this work.

E.2 PFAS FET Sensor Probe

For the first application: PFAS FET Sensor Probe. The query as shown in D.1 is:

PFAS FET Sensor Probe

```
Which two-dimensional (2D) nanomaterials or molecular probes-including novel,
↪  underexplored candidates-offer the greatest potential to achieve highly
↪  sensitive and selective detection of per- and polyfluoroalkyl substances
↪  (PFAS), or more broadly chemical and biological analytes, when integrated
↪  into a field-effect transistor (FET)-based sensor platform?
```

The best commercial solution is: **ChatGPT_o3_DR**.
The best local solution is: **DToR_qwq32B_local500**.
The expert based on the two reports, provided 3 commercial DR report derived candidates (A1-A3), and 3 local DR report derived candidates (B1-B3), visualized in DL_1 1:

A1 — 2D rGO + β-Cyclodextrin (β-CD)

Hexagonal P1 slab with a = b = 36.90 Å, γ = 120°, c = 45.00 Å (in-plane area ≈ 1179.19 Å^2). Atom counts (C/H/O = 483/102/94) indicate a 2D graphene-derived sheet with oxygenated defects (epoxy/hydroxyl) on one face and a carbohydrate torus consistent with β-CD mounted above it. The populated thickness is ≈ 15.02 Å (vacuum ≈ 29.98 Å); the 2D carbon plane sits near z ≈ 7.5 Å. The β-CD presents a toroidal pocket with a projected mouth ∼**15.3 Å diameter**, oriented normal to the basal plane for PFOS capture. Oxygen sites on rGO act as plausible anchoring points while the underlying 2D lattice remains largely sp^2 and planar.

A2 — 2D MoS$_2$/2D hBN heterojunction + methacrylic-acid (MIP-like) cavity, pyrene-linked

Hexagonal P1 with a = b = 28.44 Å, γ = 120°, c = 46.13 Å (area ≈ 700.47 Å^2). Composition shows a **2D MoS$_2$/2D hBN heterojunction** stack capped by an aromatic (pyrene-type) tether and a **methacrylic-acid cluster forming a β-CD–like cavity** on one side (organics C/H/O sit above the dielectric). The populated thickness is ≈ 18.68 Å (vacuum ≈ 27.45 Å). The cavity's projected mouth is ∼**13.6 Å in diameter**, positioned above hBN so PFOS can approach without perturbing the MoS$_2$ channel. The pyrene linker bridges π–π to the 2D cap while keeping the cavity axis roughly perpendicular to the surface, preserving electronic decoupling from the channel.

A3 — 2D Fluorographene (CF) slab

Hexagonal P1 with a = b = 36.90 Å, γ = 120°, c = 45.00 Å (area ≈ 1179.19 Å^2). Stoichiometry $C_{450}F_{450}$ yields a single-layer **2D** CF sheet with chair-like alternating C–F buckling; the "solid" thickness is ≈ 3.26 Å and vacuum ≈ 41.74 Å. The surface is uniform and free of additional adsorbates, providing a continuous fluorinated interface. Fluorine is arranged on both sides of the 2D carbon lattice, giving a symmetric fluorophilic environment. The wide lateral extent minimizes edge effects, making it suitable for baseline PFOS adsorption comparisons.

B1 — 2D Ti$_3$C$_2$O$_2$ MXene + methacrylic-acid (MIP-like) cavity

Hexagonal P1 with a = b = 30.67 Å, γ = 120°, c = 46.70 Å (area ≈ 814.63 Å^2). The slab is an O-terminated **2D Ti$_3$C$_2$** MXene bearing, on one side, a **methacrylic-acid cluster organized as a β-CD–like cavity** (additional C/H/O located at higher z). Populated thickness ≈ 16.14 Å (vacuum ≈ 30.56 Å). The cavity mouth projects ∼**13.5 Å in diameter**, perched above the O-terminated surface while the bottom MXene layer serves as a rigid support. The one-sided functionalization creates an asymmetric interface—electronegative terminations and the cavity's polar rim cooperate to stabilize PFOS headgroup interactions.

B2 — 2D Graphene + methacrylic-acid (MIP-like) cavity

Hexagonal P1 with a = b = 36.90 Å, γ = 120°, c = 45.00 Å (area ≈ 1179.19 Å^2). A 15×15 **2D** graphene sheet supports, on one face, a **methacrylic-acid cluster arranged into a β-CD–like cavity** (C/H/O entirely above the carbon plane). The populated thickness is ≈ 9.58 Å with ≈ 35.42 Å vacuum. The cavity presents a projected ∼**13.6 Å diameter** single-sided pocket aligned normal to the basal plane for PFOS engagement. The 2D sp^2 lattice remains flat and conductive, making this a clean platform for decoupling recognition from transduction.

B3 — 2D Graphene + fluorinated phospholipid monolayer

Hexagonal P1 with a = b = 36.90 Å, γ = 120°, c = 45.00 Å (area ≈ 1179.19 Å^2). Elemental makeup (C/F/H/O with trace N and P) is consistent with phosphatidylcholine headgroups near the **2D** graphene plane and perfluorinated tails extending outward to form a dense brush. The populated thickness is ≈ 14.84 Å (vacuum ≈ 30.16 Å), with F spanning roughly z ≈ 9.21–22.34 Å above the basal plane. The assembly yields a hydrophilic-to-fluorophilic gradient, presenting a fluorinated outer surface for PFOS tail affinity. Lipid tails show a collective tilt relative to the surface normal, producing a compact, low-defect fluorinated interface on one side while the opposite side remains bare graphene.

Selectivity Scoring for PFAS Sensor Performance The practical deployment of PFAS sensors demands not only strong binding to target per- and polyfluoroalkyl substances but also selective discrimination against common environmental interferents. To quantitatively evaluate this selectivity, we developed a unified scoring framework applicable to both static DFT calculations and explicit-solvent AIMD simulations, enabling direct comparison between computational methodologies while maintaining thermodynamic rigor.

Selectivity Metric Derivation: The fundamental selectivity metric $\Delta\Delta E$ quantifies the differential binding preference between PFAS compounds and interferent molecules:

$$\Delta\Delta E = \overline{\Delta E}_{\text{PFAS}} - \overline{\Delta E}_{\text{Interf}} \tag{3}$$

where $\overline{\Delta E}_{\text{PFAS}}$ represents the average binding energy across four representative PFAS compounds (PFOA, PFOS, PFNA, and PFHxS), and $\overline{\Delta E}_{\text{Interf}}$ denotes the average binding energy for interferent species (humic acid and SDS), as illustrated in Figure DL_1 2. A larger positive $\Delta\Delta E$ value indicates superior selectivity, signifying stronger preferential binding to PFAS over interferents.

For static DFT calculations, ΔE values correspond to the conventional binding energies ΔE_{bind} computed under implicit solvent conditions (Figure DL_1 3). In explicit-solvent AIMD simulations, these are replaced by the time-averaged end-state potential energy gaps $\Delta E^{\text{ES}}_{\text{pot}}$, which capture solvent reorganization effects and hydrogen-bonding network rearrangements absent from static calculations (Figures DL_1 4 and DL_1 5).

Linear Selectivity Scoring Framework: To enable meaningful comparison across candidate systems and computational methods, we normalize the selectivity metrics to a common scoring scale:

$$\text{Score} = 100 \times (1 + \beta(\Delta\Delta E - \Delta\Delta E_{\beta\text{-CD}})) \tag{4}$$

where $\Delta\Delta E_{\beta\text{-CD}}$ represents the selectivity of β-cyclodextrin, our domain benchmark reference with established PFAS binding capability [23]. The sensitivity parameter $\beta = 0.2$ was selected to provide appropriate score discrimination while maintaining numerical stability across both DFT and AIMD datasets. This linear transformation preserves the physical meaning of energy differences while establishing β-CD as the baseline reference (Score = 100).

Comparative Analysis of Selectivity Under Different Solvation Models: The application of this unified scoring framework reveals striking differences between implicit and explicit solvation treatments. In static DFT calculations with implicit solvent, system A1 achieves the highest selectivity score of 102.4, outstandingly exceeding the β-CD baseline (Figure DL_1 6). The B-series candidates also prove competitive as B2 and B3 also showed slightly passing 100 score.

However, the inclusion of explicit water molecules and dynamic sampling in AIMD simulations dramatically alters the selectivity landscape (Figure DL_1 7). Most notably, all candidate systems score below the β-CD baseline in explicit solvent, with B1 achieving the highest relative performance at 95.7. This systematic reduction in scores reflects the critical role of solvent-mediated effects: competitive water adsorption at binding sites, strong hydration of anionic PFAS headgroups, and the entropic penalties associated with constraining flexible molecules in adsorbed configurations.

The reversal of relative rankings between static DFT and AIMD calculations—particularly the superior performance of B-series candidates under explicit solvation—underscores the importance of including solvent dynamics in sensor design evaluation. While static calculations provide valuable initial screening, the explicit-solvent AIMD results more accurately capture the complex interplay between adsorbate–surface interactions and solvent reorganization that governs selectivity under realistic aqueous conditions.

Implications for Sensor Design: The divergence between implicit and explicit solvent results carries significant implications for PFAS sensor development. Systems optimized solely based on static DFT calculations may exhibit diminished selectivity in aqueous environments due to unforeseen solvent competition effects. Conversely, candidates that appear suboptimal in vacuum or implicit solvent calculations may benefit from favorable solvent-mediated interactions that enhance their practical selectivity. This highlights the necessity of multi-scale computational approaches that progress from efficient static screening to rigorous explicit-solvent validation for promising candidates.

The linear selectivity scoring framework presented here provides a quantitative, thermodynamically grounded metric for evaluating sensor performance across computational methodologies. By maintaining consistent notation and normalization across static and dynamic simulations, this approach enables systematic comparison of candidate materials while preserving the physical interpretability of the underlying energy differences.

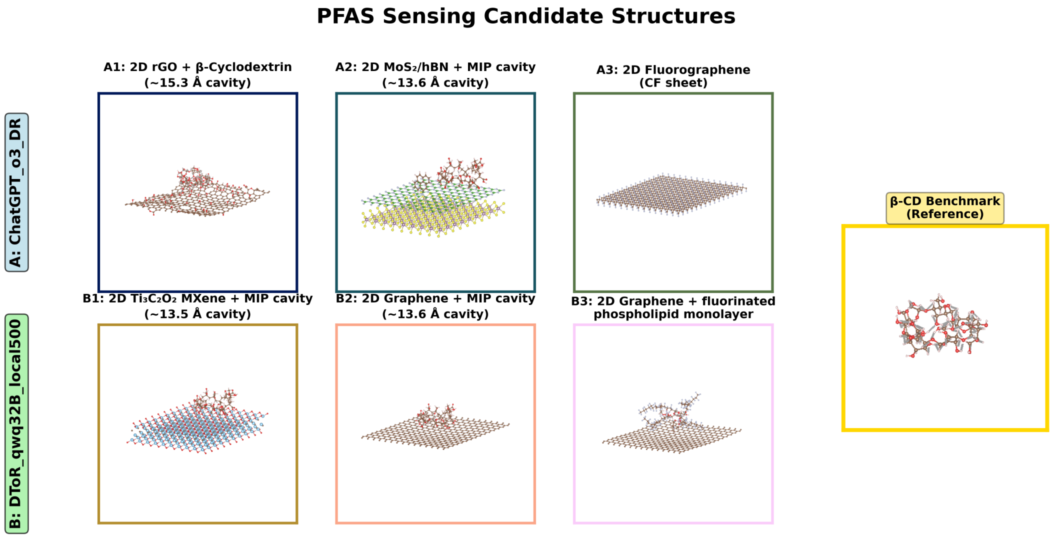


Figure DL_1 1: Structural configurations of candidates proposed by DR agents for selective PFAS binding, with β-cyclodextrin [23] included as the domain benchmark reference.

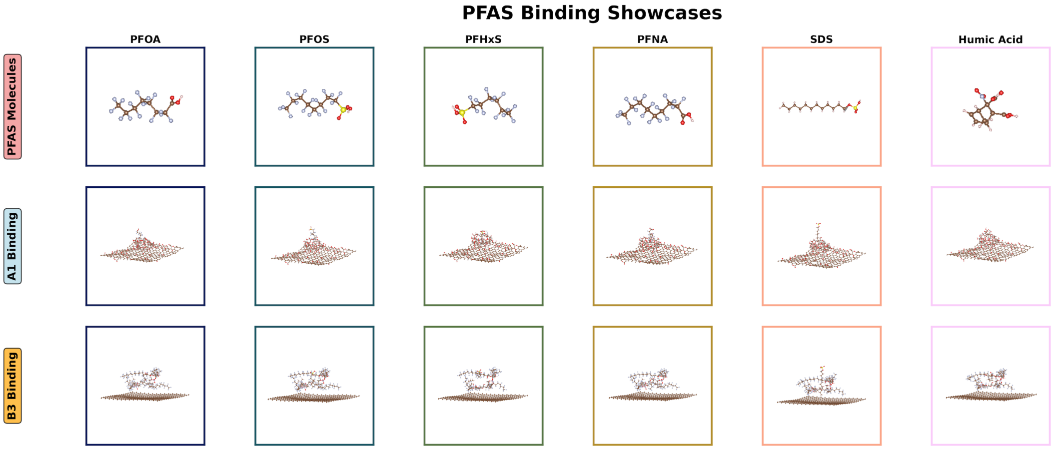


Figure DL_1 2: Representative PFAS compounds (PFOA, PFOS, PFNA, PFHxS) and interferent molecules (SDS and humic acid) employed in selectivity assessment, with corresponding optimized binding configurations. Two candidates' relaxed structures are shown.

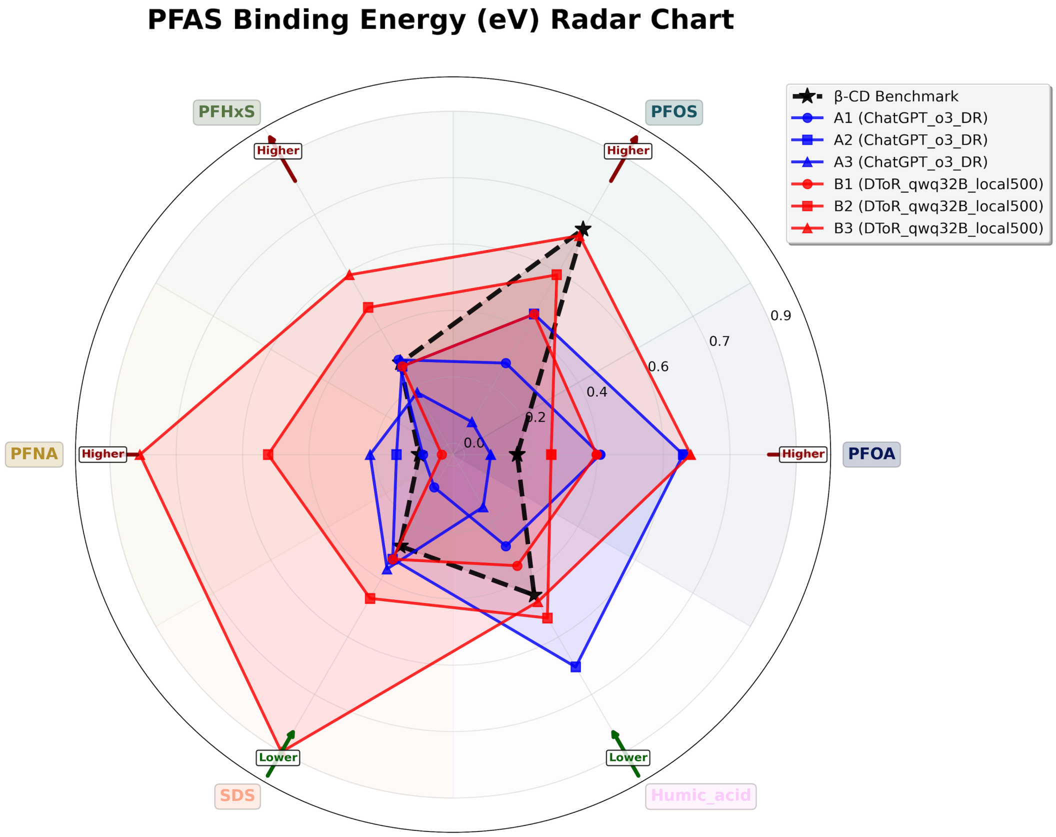


Figure DL_1 3: Radar plot comparing binding energies (ΔE_{bind}) across all candidate systems and benchmark β-CD obtained from static DFT calculations with implicit solvent.

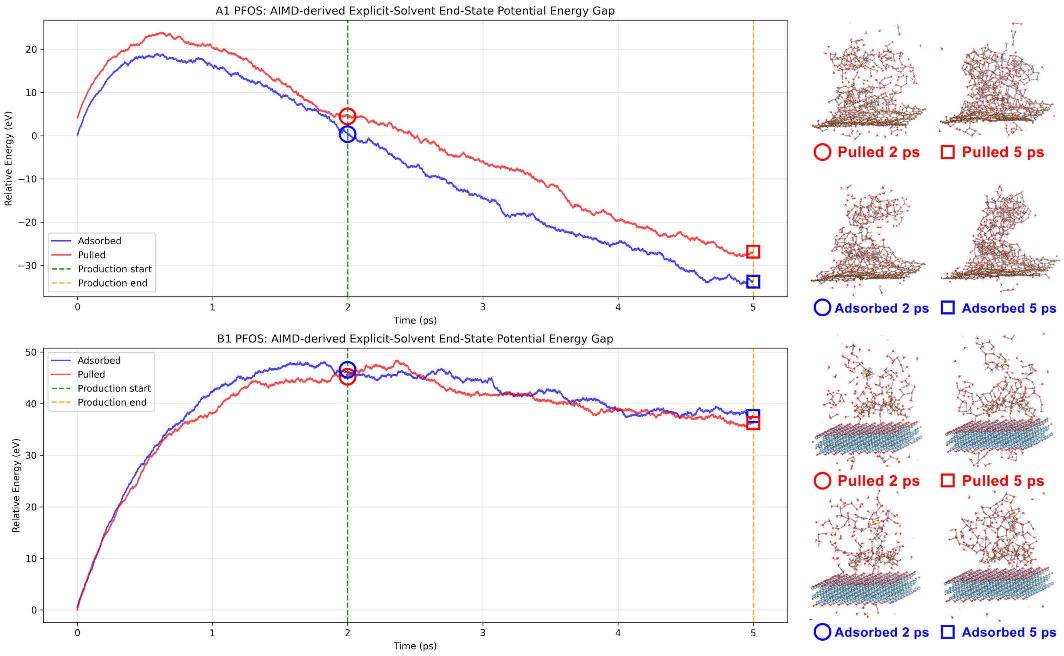


Figure DL_1 4: Representative AIMD trajectory showing relative energy evolution under explicit solvent conditions. The green dashed line denotes the 2 ps equilibration period. Right panels display system configurations at highlighted time points along the production trajectory.

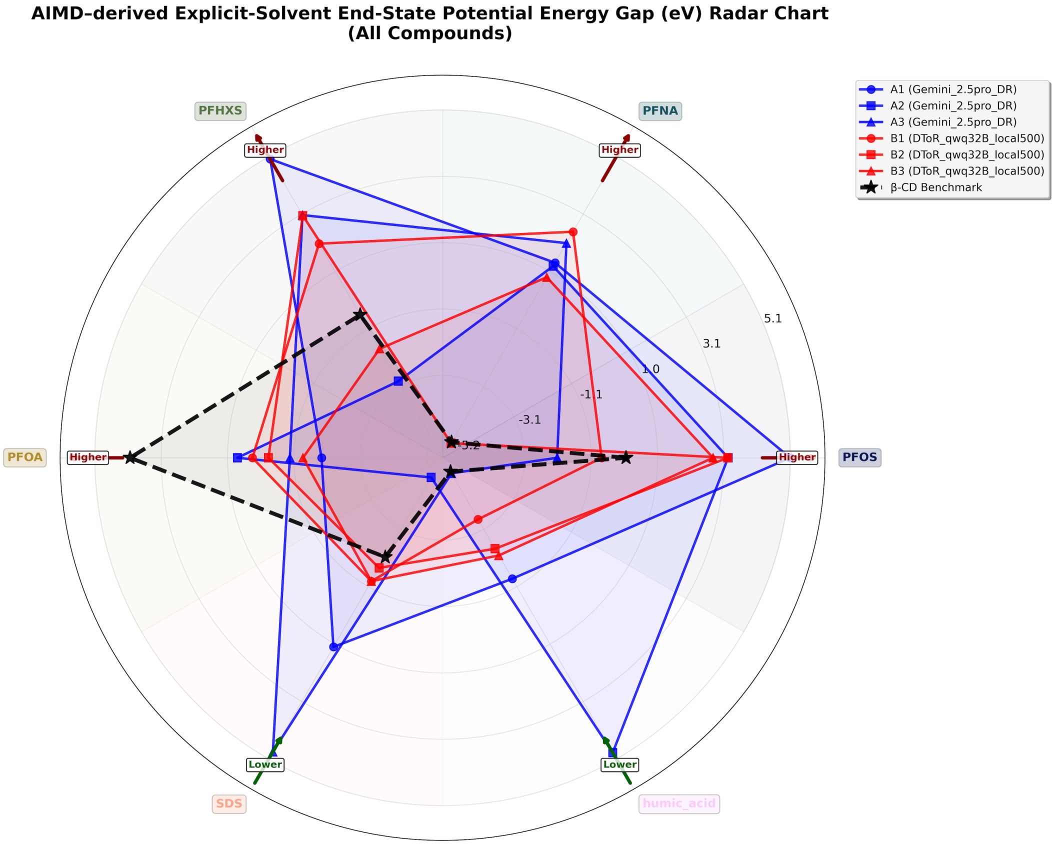


Figure DL_1 5: Radar plot comparing explicit-solvent end-state potential energy gaps ($\Delta E_{\text{pot}}^{\text{ES}}$) averaged over the 3 ps production period for all candidate systems and benchmark β-CD from AIMD simulations.

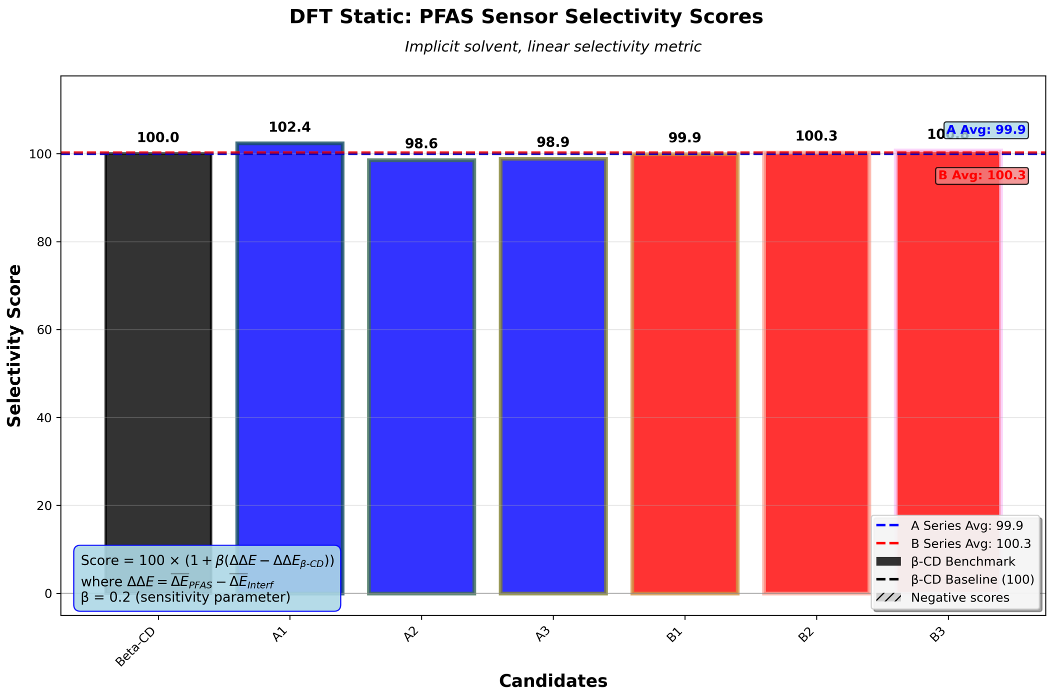


Figure DL_1 6: Selectivity scores for all candidate systems from static DFT calculations, normalized to β-CD baseline (100). Scores are calculated using the linear selectivity metric with sensitivity parameter $\beta = 0.2$, where $\Delta\Delta E = \overline{\Delta E}_{\text{PFAS}} - \overline{\Delta E}_{\text{Interf}}$ represents the difference between average PFAS and interferent binding energies.

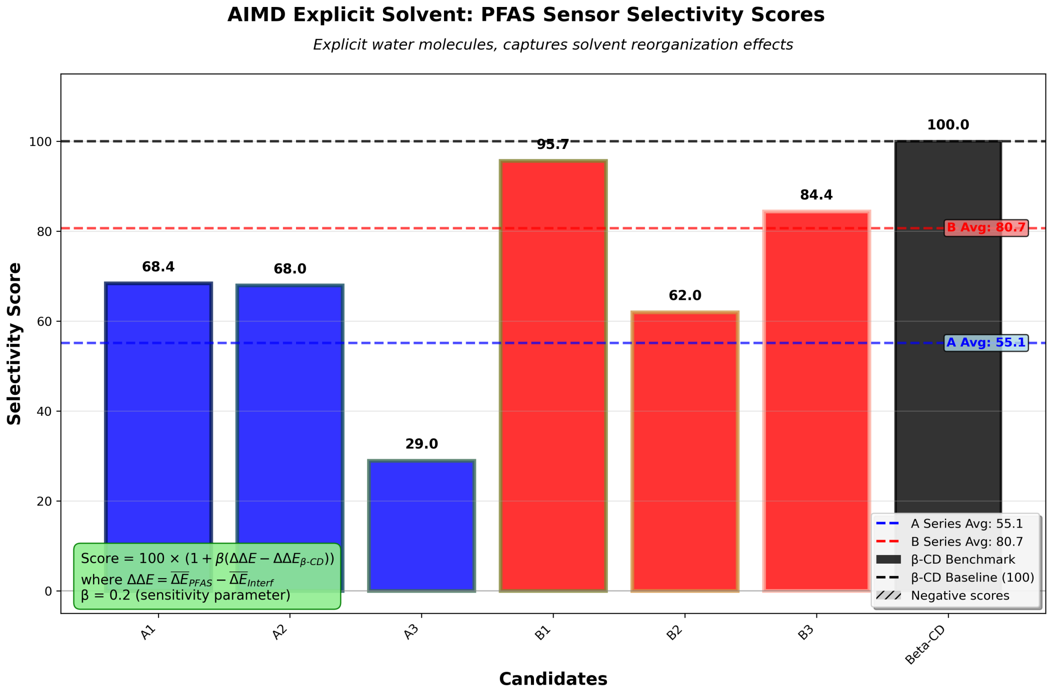


Figure DL_1 7: Selectivity scores for all candidate systems from explicit-solvent AIMD simulations, normalized to β-CD baseline (100). The scoring methodology remains identical to static DFT, except ΔE values are replaced by $\Delta E^{\text{ES}}_{\text{pot}}$, the time-averaged potential energy gaps from the 3 ps production period.

E.3 PFAS Degradation

For the second application: PFAS Degradation. The query as shown in D.1 is:

PFAS Degradation

```
Which novel electrode materials can achieve efficient PFAS degradation under
↪  ambient aqueous electrochemical conditions, delivering both high
↪  mineralization and defluorination rates? What intrinsic properties-such as
↪  PFAS adsorption affinity, reactive oxygen species generation capacity, and
↪  C-F bond activation energy-should be optimized to guide their discovery?
```

The best commercial solution is: **Gemini_2.5_Pro_DR**.

The best local solution is: **DToR_qwq32B_local500**.

The expert based on the two reports, provided 5 commercial DR report derived candidates (A1-A5), and 5 local DR report derived candidates (B1-B5), visualized in DL_2 1:

Ti_4O_7 — Pristine Magnéli (001) bench slab

Triclinic P1 slab with a = 22.376 Å, b = 27.651 Å, $\gamma = 75.306°$, c = 42.208 Å (in-plane area ≈ 598.50 Å^2). Atom counts (Ti/O = 256/448; 704 total) match a Ti_4O_7 (001) termination. The populated thickness spans ≈ 12.13 Å (z ≈ 3.41–15.53 Å), leaving ≈ 30.08 Å vacuum; the slab mid-plane lies near z ≈ 9.47 Å. Surfaces are stoichiometric and flat, preserving Magnéli shear-plane connectivity and metallic conduction pathways—our baseline conductive-oxide reference for PFAS tests. Build/relax notes: 4 atomic layers; bottom two constrained. (Adsorbates for subsequent runs were seeded ≈ 3 Å above the most protruding Ti site, tail parallel to the surface.)

A1 — Ti_4O_7 (001) with through-pore + rim O-vacancies

Triclinic P1 cell with a = 22.376 Å, b = 27.651 Å, $\gamma = 75.306°$, c = 42.208 Å (in-plane area ≈ 598.50 Å^2). Atom counts (Ti/O = 249/437; 686 total) reflect removal of a top-side cluster and additional rim O-vacancies relative to the bench ($\Delta = -7$ Ti, -11 O). The populated thickness is ≈ 13.44 Å (z ≈ 3.00–16.45 Å), leaving ≈ 28.77 Å vacuum; the slab mid-plane sits near z ≈ 9.73 Å. A circular through-pore was introduced by deleting atoms within a ∼5 Å-radius cylinder from the upper layers, yielding an effective mouth ∼1.0 nm diameter; the rim exposes under-coordinated Ti with a locally reduced O environment, while the surrounding lattice retains Magnéli shear-plane connectivity. Relaxation protocol used: bottom half constrained; top region (and any adsorbates) relaxed so the pore edge dominates binding/activation. Connection to the Ti_4O_7 bench: same lattice, cell vectors, and basal (001) termination as the pristine Ti_4O_7 bench slab; bottom region unchanged. All edits are localized to the top surface (pore machining + rim O-vacancies). Away from the pore, Ti/O stoichiometry per layer and shear-plane topology match the bench reference.

A2 — Pd-overlayer/Ti_4O_7 (001) composite (amorphous Pd on the bench slab)

Triclinic P1 cell with a = 22.376 Å, b = 27.651 Å, $\gamma = 75.306°$, c = 42.21 Å (in-plane area ≈ 598.50 Å^2). Atom counts (Ti/O/Pd = 256/448/94; 798 total) indicate a stoichiometric Ti_4O_7 substrate capped by a Pd overlayer. The populated thickness spans ≈ 16.45 Å (z ≈ 3.14–19.59 Å), leaving ≈ 25.76 Å vacuum; mid-plane at z ≈ 11.37 Å. Pd atoms populate the top terrace broadly (no single island), forming a disordered sheet with Pd heights ∼16.12–19.59 Å (mean ≈ 17.71 Å), i.e., ∼3.1 Å above the top oxide layer ($z_{ox,max} \approx 16.49$ Å). **Connection to the Ti_4O_7 bench:** identical lattice and Ti/O stoichiometry to the pristine Magnéli (001) slab; modifications are limited to a Pd overlayer on the top surface (bottom half retained as in the bench), with no pores or O-vacancies introduced.

A3 — Pd$_{55}$ cluster on Ti$_4$O$_7$ (001) (icosahedral-like nanoparticle on the bench slab)

Triclinic P1 cell with a = 22.376 Å, b = 27.651 Å, γ = 75.306°, c = 42.21 Å (in-plane area ≈ 598.50 Å^2). Atom counts (Ti/O/Pd = 256/448/55; 759 total) show a Ti$_4$O$_7$ substrate supporting a compact Pd nanoparticle. The populated thickness spans ≈ 22.65 Å (z ≈ 3.23–25.87 Å), leaving ≈ 19.56 Å vacuum; mid-plane at z ≈ 14.55 Å. The Pd cluster (initially icosahedral; allowed to relax) (n = 55) is localized on the top terrace with a projected footprint ∼9.2 × 9.6 Å (max in-plane radius ≈ 5.33 Å; projected diameter ≈ **10.65 Å**) and vertical extent placing the apex ∼9.7 Å above the top oxide layer ($z_{\mathrm{ox,max}}$ ≈ 16.15 Å). **Connection to the Ti$_4$O$_7$ bench:** same lattice and stoichiometric oxide as the pristine (001) slab; the only alteration is a single Pd$_{55}$ nanoparticle anchored on the top surface (no pore machining or O-vacancies), with the bottom region kept as in the bench reference.

A4 — F-doped rutile SnO$_2$ (110) slab (FTO, O→F substitution near surface)

Orthogonal P1 cell with a = 26.8022 Å, b = 25.4880 Å, γ = 90.0°, c = 42.1012 Å; in-plane area ≈ 683.13 Å^2. The metrics correspond to a 4×8 SnO$_2$(110) supercell ($a \approx 4\times a_{(\mathrm{rutile})}\sqrt{2}$, $b \approx 8\times c_{(\mathrm{rutile})}$). Atom counts (Sn/O/F = 256/497/15; 768 total) indicate ∼2.9% O→F substitution on the anion sublattice (15/512), i.e., a lightly F-doped "FTO" slab. The populated thickness spans ≈ 14.28 Å (z ≈ 4.32–18.60 Å), leaving ≈ 27.83 Å vacuum; Sn layers occupy z ≈ 4.38–15.29 Å, while F sites are confined to a narrow band near the top (z ≈ 16.33–17.17 Å; mean ≈ 16.75 Å), just beneath the outermost O ridges (O_{max} ≈ 18.60 Å). This places F primarily in near-surface O positions on (110), preserving the characteristic bridging-O/in-plane-O relief of rutile and providing a slightly more electron-rich, less hydroxylated top termination expected for FTO.

A5 — B-doped four-layer graphene

Hexagonal P1; a = b = 19.68 Å, γ = 120°, **c ≈ 36.0 Å**. Four graphene sheets stacked with the standard **interlayer spacing d(002) ≈ 3.35 ± 0.05 Å** (three gaps → ∼10.05 Å slab thickness). Substitutional B is confined to the **top layer** (≈10 at% within that plane; ≈2–3 at% overall), preserving an sp^2 framework. The remaining **∼26 Å is vacuum** along c to suppress image interactions. During relaxation, we fixed the bottom two layers; relax the top two (and any adsorbates).

B1 — F-grafted four-layer graphene (consensus spacing; A5-consistent)

Hexagonal P1; **keep A5's in-plane metrics (a = b, γ = 120°) unchanged** and **c ≈ 36.0 Å**. Stack **four graphene sheets with the standard interlayer spacing d(002) = 3.35 ± 0.05 Å** (three gaps → ∼10.05 Å slab thickness), with the **remaining ∼26 Å as vacuum** along c to decouple periodic images. **Chemistry/dopant pattern exactly as in A5** (no changes to which layer(s) are doped/functionalized); the carbon network remains predominantly sp^2. **Same Relaxation protocol as A5:** fix the bottom two layers; relax the top two (and any adsorbates). **Connection to A5:** identical spacing convention and vacuum budget as A5 to ensure cross-comparable adsorption energies and electrostatics across the graphene series, while leaving only the dopant/functional differences to drive behavior.

B2 — 2D Ti_3C_2 MXene (O/OH-terminated)

Hexagonal P1 slab with a = b = 24.536 Å, γ = 120°, c = 36.700 Å (in-plane area ≈ 521.36 $Å^2$). Atom counts (Ti/C/O/H = 192/128/128/114; 562 total) correspond to **64 f.u. of Ti_3C_2** with full two-sided terminations (128/site total): ≈**89% –OH** (114 H) and ≈**11% =O** (14 O), distributed symmetrically per face (**57 –OH + 7 =O** each side). The vacuum layer is around **27.21 Å**. The sheet preserves the characteristic **2D MXene** core (Ti–C–Ti trilayers) with mixed O/OH terminations pointing outward, yielding a hydrophilic, electronically conductive surface appropriate for adsorption tests (e.g., PFAS headgroup coordination via –OH domains).

B3 — 2D MnO_2-like (birnessite-style) nanosheet with sparse one-sided –OH

Orthogonal P1 cell with a = b = 29.3628 Å, γ = 90.0°, c = 33.595 Å (in-plane area ≈ 862.17 $Å^2$). Atom counts (Mn/O/H = 216/411/18; 645 total) are consistent with a near-stoichiometric MnO_2 layer carrying light hydroxylation (H:Mn ≈ 0.083). The populated thickness is ≈ 9.32 Å (z ≈ 0.82–10.14 Å), leaving ≈ 24.28 Å vacuum. Element-resolved spans: Mn ≈ 1.52–9.59 Å; O ≈ 0.82–9.61 Å; H confined to ≈ 7.12–10.14 Å on the outer face, indicating **one-sided –OH/water terminations** on a **2D** octahedral sheet. Relaxation note: we kept the lower half constrained and relaxed the top half/adsorbates; head-group binding is expected at surface –OH/O sites with the basal Mn–O framework remaining robust.

B4 — 2D MnO_2-like (birnessite-style) nanosheet, smaller supercell; one-sided –OH

Orthogonal P1 cell with a = b = 22.000 Å, γ = 90.0°, c = 33.610 Å (in-plane area = 484.00 $Å^2$). Composition (Mn/O/H = 150/285/12; 447 total) again reflects a MnO_2 sheet with light hydroxylation (H:Mn ≈ 0.08). The populated thickness is ≈ 9.40 Å (z ≈ 1.77–11.17 Å), leaving ≈ 24.21 Å vacuum. Element-resolved spans: Mn ≈ 2.12–10.48 Å; O ≈ 1.77–10.99 Å; H ≈ 8.43–11.17 Å, i.e., **–OH/water groups localized on the top face** of the **2D** sheet. Relaxation matches B3 (bottom region constrained, top free) to keep comparisons clean across the MnO_2 series.

B5 — Graphene-capped 2D MXene (derived from B2)

Derived directly from **B2** (same cell vectors and terminations), a **single-layer graphene** sheet is added parallel to the top face of the MXene and centered to span the full in-plane cell. The initial MXene$\leftrightarrow$graphene separation is set to $\sim$**3.35 Å** (graphite d(002)), with **no covalent links**—pure van der Waals contact. Graphene is pristine sp^2 and allowed to relax laterally/vertically; the MXene beneath follows B2's relaxation protocol (bottom constrained, top free). The **c-axis vacuum is kept as in B2** to avoid inter-image coupling. **Connection to B2:** identical substrate; the only modification is the graphene capping layer.

Multi-Metric Evaluation of PFAS Degradation Catalysts

Effective catalytic degradation of PFAS compounds requires both strong surface adsorption to concentrate the pollutants and sufficient activation of the exceptionally stable C–F bonds. To comprehensively evaluate candidate catalysts for PFAS degradation, we employed a three-pronged computational assessment framework encompassing static binding energies, explicit-solvent dynamics, and C–F bond dissociation energetics, with Ti_4O_7 serving as the domain benchmark reference [49].

Integrated Performance Metrics

The degradation performance evaluation integrates three complementary metrics that capture distinct aspects of the catalytic process. First, the binding energy ΔE_{bind} from static DFT calculations quantifies the thermodynamic driving force for PFAS adsorption onto catalyst surfaces (Figure DL_2 3). Strong binding ensures sufficient surface coverage and residence time for subsequent degradation reactions. Second, the explicit-solvent potential energy gap $\Delta E_{\text{pot}}^{\text{ES}}$ from AIMD simulations reveals how aqueous solvation affects the adsorption thermodynamics (Figures DL_2 7 and DL_2 8), accounting for competitive water adsorption and hydrogen-bonding network reorganization that static calculations cannot capture. Third, the C–F bond dissociation energy quantifies the activation barrier for the rate-limiting defluorination step, computed by displacing fluorine atoms from adsorbed PFAS configurations to isolated positions within the supercell (Figure DL_2 4).

Scoring Framework for Degradation Performance

To enable quantitative comparison across these diverse metrics, we applied a unified linear scoring framework:

$$\text{Score} = 100 \times \left(1 + \beta(\Delta\bar{E} - \Delta\bar{E}_{\mathrm{Ti_4O_7}})\right) \tag{5}$$

where $\Delta\bar{E}$ represents the mean energy value averaged over four representative PFAS compounds (PFOA, PFOS, PFNA, PFHxS), and $\beta = 0.2$ serves as the sensitivity parameter. For binding metrics (both static DFT and AIMD), more negative $\Delta\bar{E}$ values indicate stronger adsorption and thus higher scores. Conversely, for C–F dissociation, lower energy barriers (less positive values) yield higher scores, reflecting more facile bond activation.

The distribution of C–F dissociation energies across different PFAS compounds and candidate surfaces exhibits substantial variability (Figure DL_2 5), with mean values ranging from approximately 2 to 6 eV (Figure DL_2 6). This variation underscores the importance of averaging over multiple PFAS species to obtain representative performance metrics, as individual compounds may exhibit preferential interactions with specific surface sites or crystallographic facets.

Performance Analysis Across Computational Methods

Application of this scoring framework reveals distinct performance patterns across the three evaluation metrics. In static DFT calculations, several candidates demonstrate superior PFAS binding relative to the Ti_4O_7 benchmark (Figure DL_2 9), with A-series materials generally exhibiting stronger adsorption energies. However, the inclusion of explicit solvent effects in AIMD simulations significantly alters these rankings (Figure DL_2 11), with most candidates showing reduced binding scores due to solvent-mediated destabilization effects. This divergence highlights the critical importance of accounting for aqueous solvation in evaluating catalyst performance under realistic conditions.

The C–F bond activation scores (Figure DL_2 10) provide complementary insights into degradation potential. Notably, candidates that excel in binding may not necessarily facilitate efficient C–F bond cleavage, as these properties depend on distinct electronic and geometric factors. The ideal degradation catalyst must balance strong PFAS adsorption with effective bond activation—a trade-off that our multi-metric scoring framework quantitatively captures.

Implications for Catalyst Design

The comprehensive evaluation framework presented here reveals that no single candidate uniformly excels across all performance metrics. Strong surface binding, while essential for PFAS capture, does not guarantee facile C–F bond activation. Similarly, materials optimized for bond dissociation may suffer from weak adsorption, particularly under aqueous conditions where solvent competition is significant. This multi-dimensional performance landscape emphasizes the need for balanced catalyst design that considers both thermodynamic (binding) and kinetic (bond activation) factors, evaluated under both implicit and explicit solvation conditions to ensure robust performance predictions for practical applications.

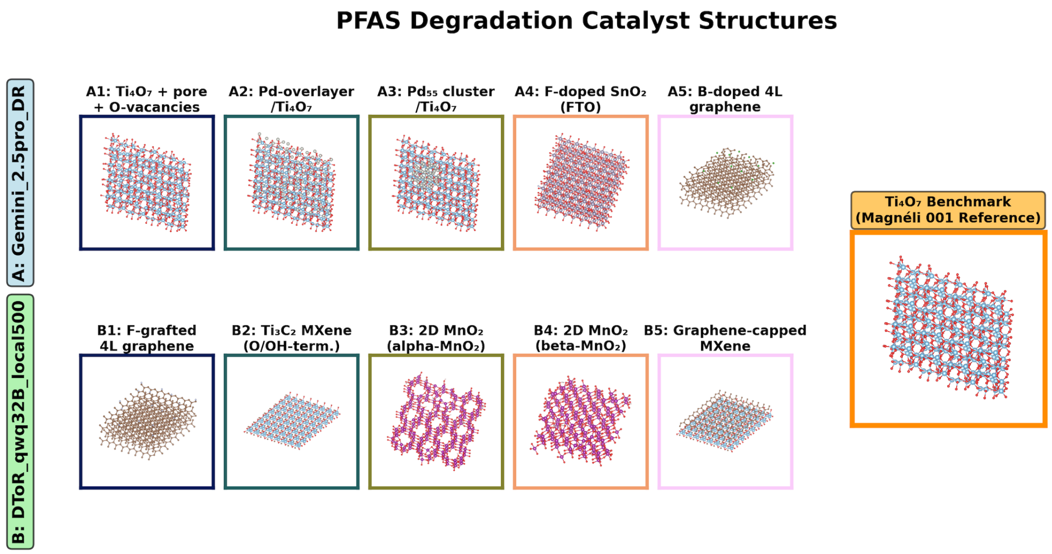


Figure DL_2 1: Structural configurations of candidate catalysts proposed by DR agents for PFAS degradation, with Ti_4O_7 [49] included as the domain benchmark reference.

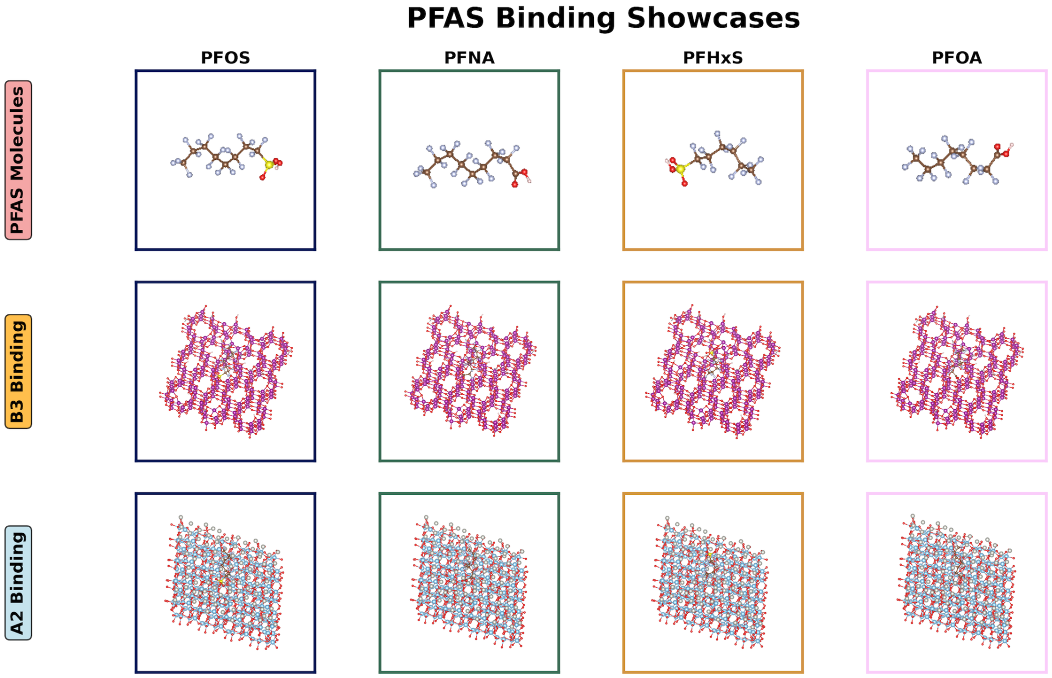


Figure DL_2 2: Representative PFAS compounds (PFOA, PFOS, PFNA, PFHxS) employed in degradation assessment, with corresponding optimized adsorption configurations on selected candidate surfaces.

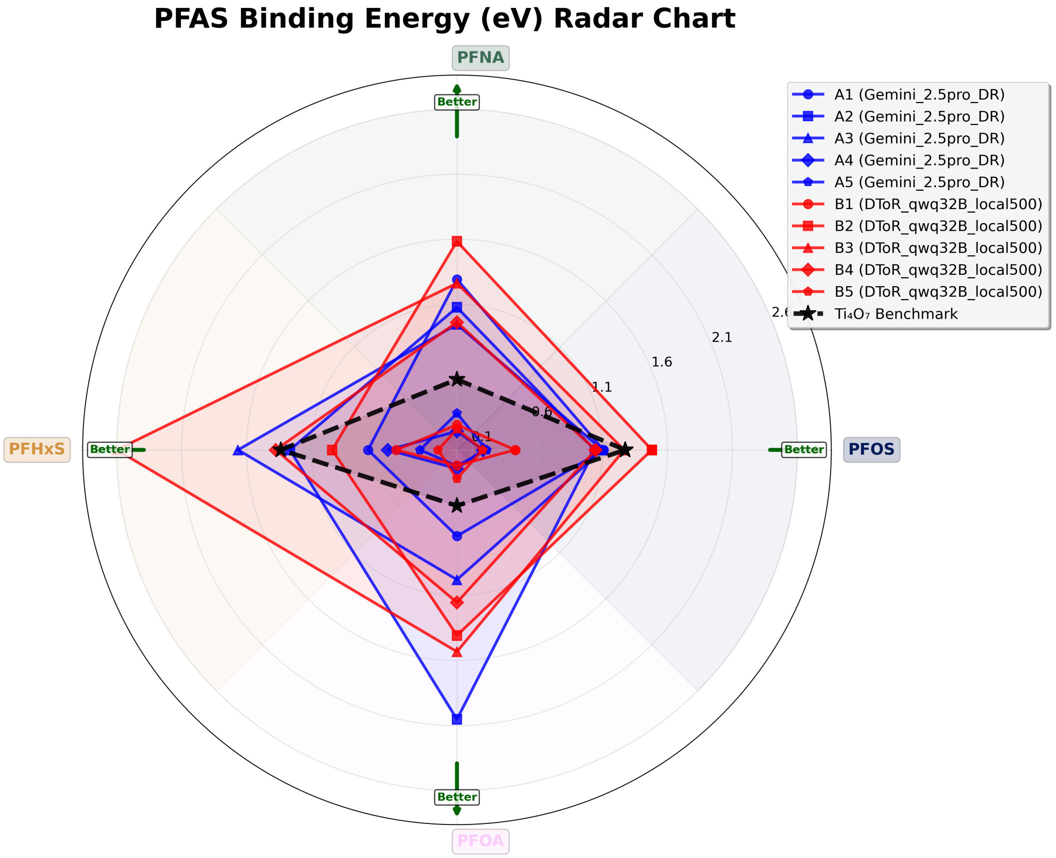


Figure DL_2 3: Radar plot comparing PFAS binding energies (ΔE_{bind}) across all candidate catalysts and benchmark Ti_4O_7 obtained from static DFT calculations with implicit solvent.

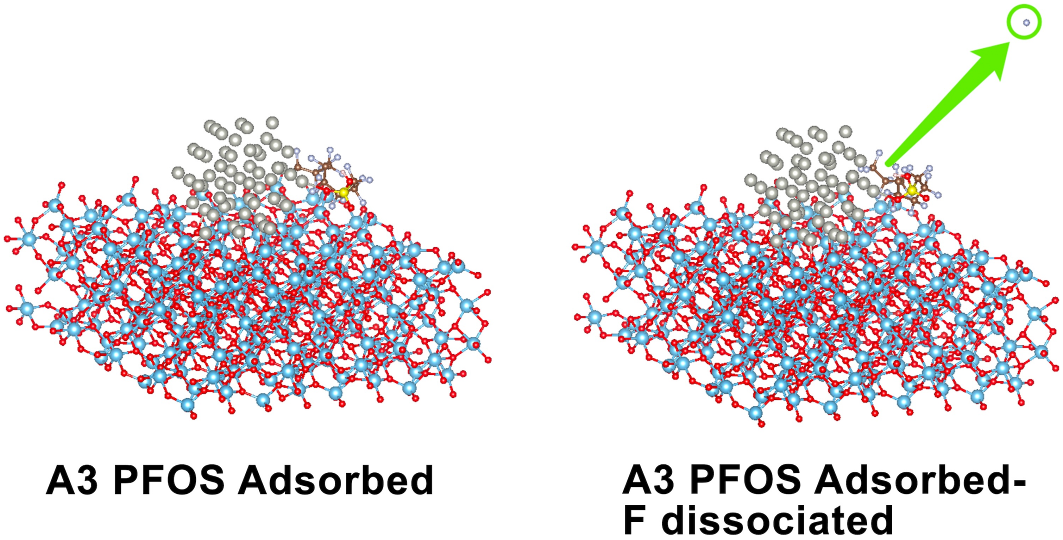


Figure DL_2 4: Illustration of C–F bond dissociation simulation protocol. Left: Initial configuration with PFOS adsorbed on A3 surface. Right: Final configuration with fluorine atom displaced to corner position (5 Å from each boundary) of the supercell.

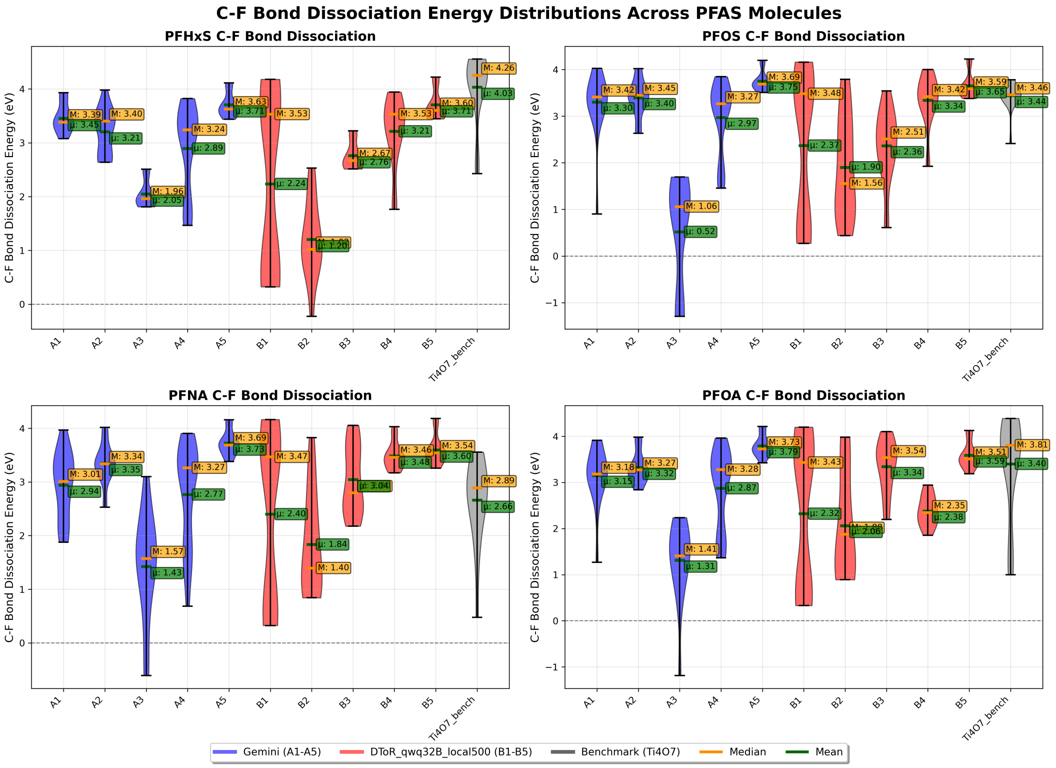


Figure DL_2 5: Violin plots showing distributions of C–F bond dissociation energies for different PFAS compounds (PFOA, PFOS, PFNA, PFHxS) across all candidate catalysts, with mean values indicated by horizontal bars.

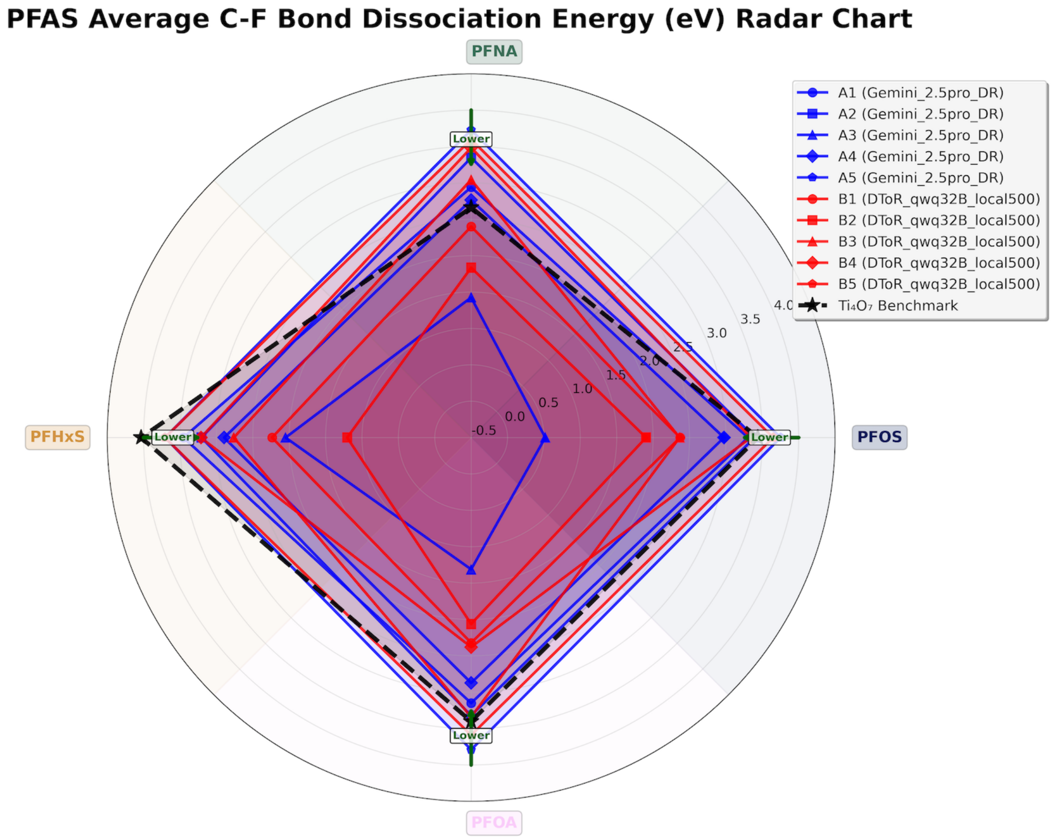


Figure DL_2 6: Radar plot comparing mean C–F bond dissociation energies across all candidate catalysts and benchmark Ti_4O_7. Lower dissociation energies indicate more favorable C–F bond activation for degradation.

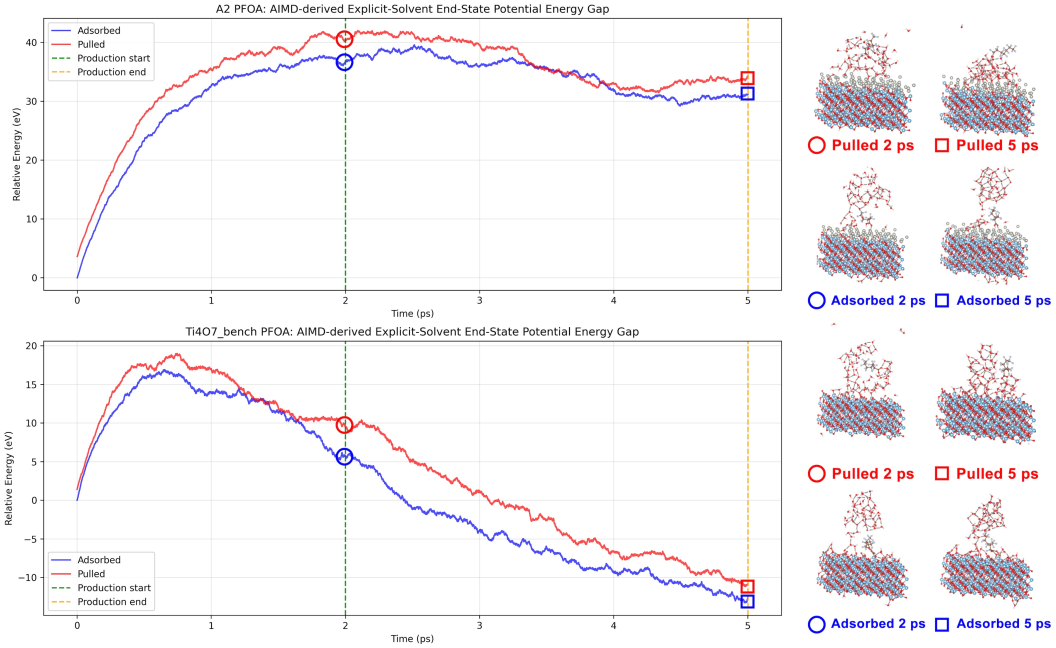


Figure DL_2 7: Representative AIMD trajectory showing system energy evolution under explicit solvent conditions. The green dashed line denotes the 2 ps equilibration period. Right panels display system configurations at highlighted time points along the 3 ps production trajectory.

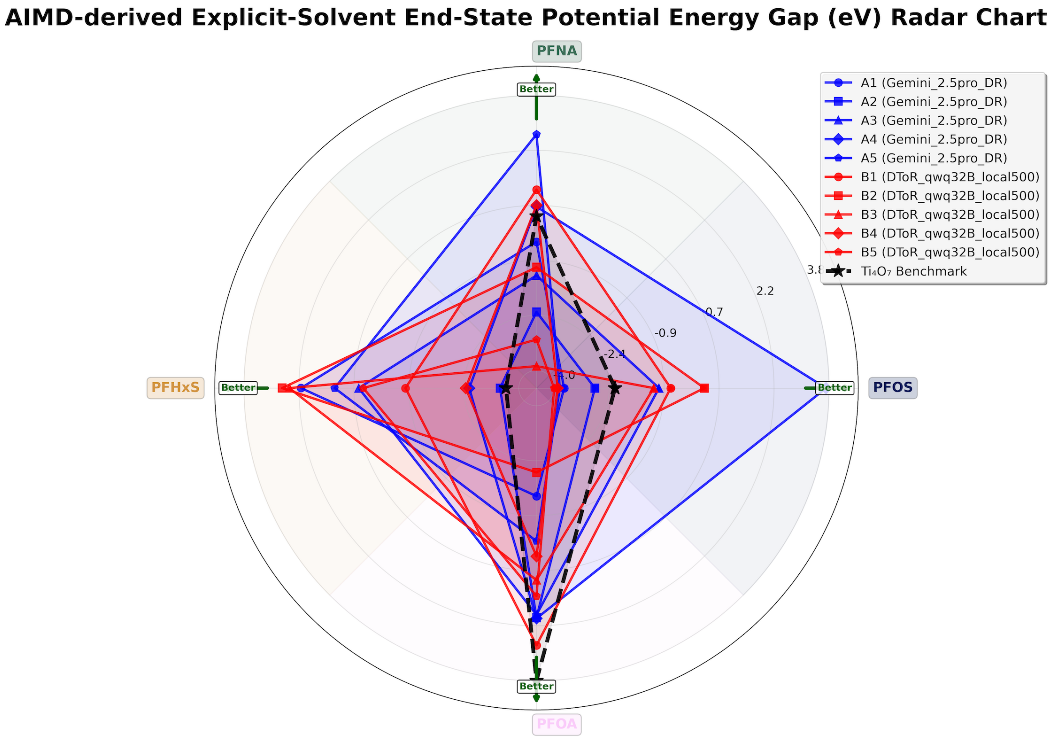


Figure DL_2 8: Radar plot comparing explicit-solvent end-state potential energy gaps ($\Delta E_{\text{pot}}^{\text{ES}}$) averaged over the 3 ps production period for all candidate catalysts and benchmark Ti_4O_7 from AIMD simulations.

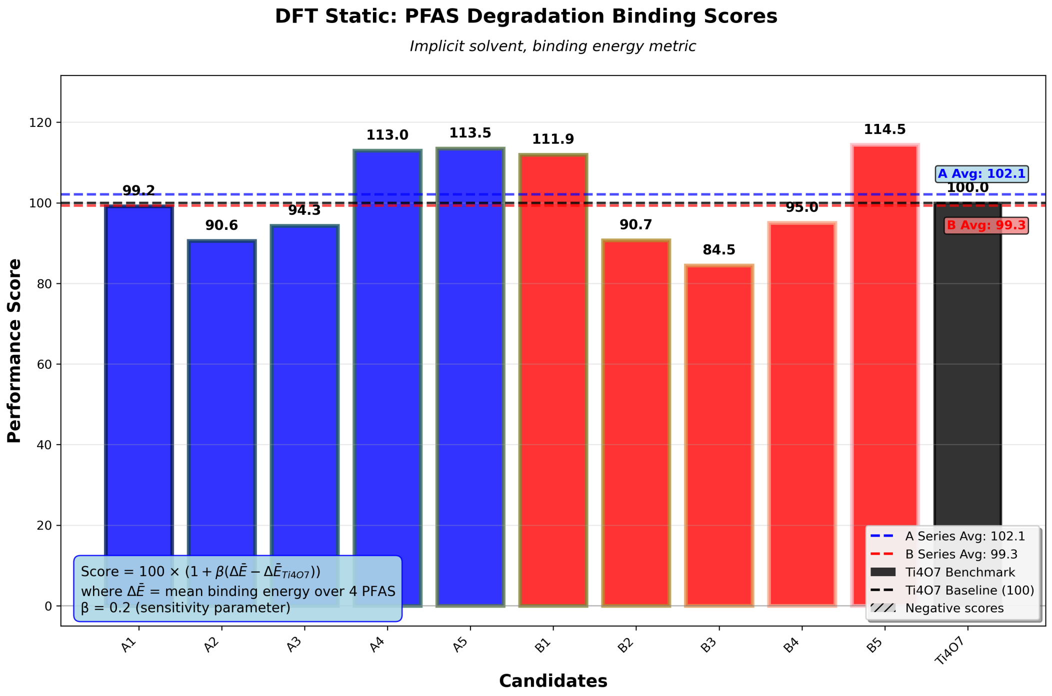


Figure DL_2 9: PFAS binding scores for all candidate catalysts from static DFT calculations, normalized to Ti_4O_7 baseline (100). Scores are calculated using the linear metric with sensitivity parameter $\beta = 0.2$, where $\Delta\bar{E} = \overline{\Delta E}_{\text{PFAS}}$ represents the mean binding energy averaged over four PFAS compounds.

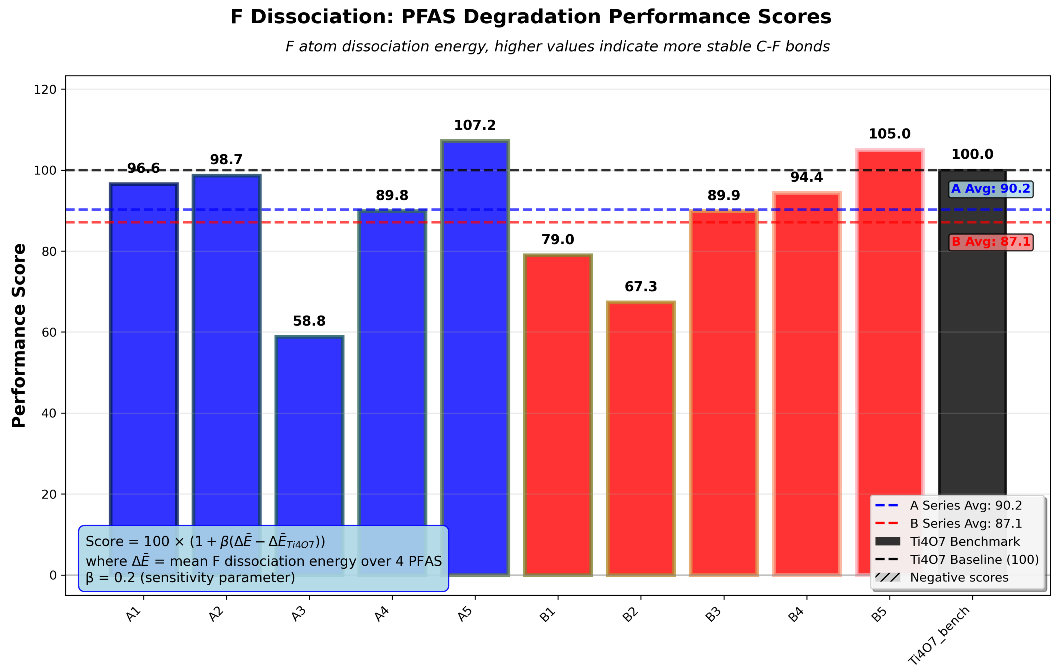


Figure DL_2 10: C–F bond activation scores for all candidate catalysts from static DFT calculations, normalized to Ti_4O_7 baseline (100). Scores are calculated using the linear metric with sensitivity parameter $\beta = 0.2$, where $\Delta\bar{E} = \overline{E}_{\text{F-dis}}$ represents the mean C–F dissociation energy averaged over four PFAS compounds. Higher scores indicate lower dissociation barriers and more favorable degradation potential.

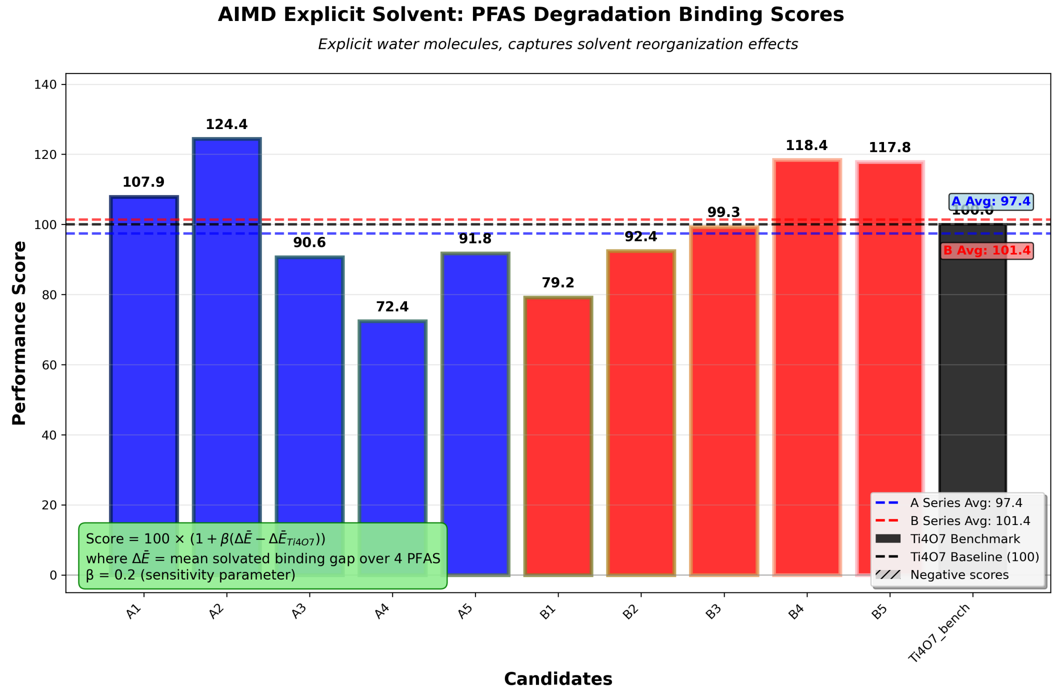


Figure DL_2 11: PFAS binding scores for all candidate catalysts from explicit-solvent AIMD simulations, normalized to Ti_4O_7 baseline (100). The scoring methodology remains identical to static DFT, with $\Delta\bar{E}$ representing the mean $\Delta E_{\text{pot}}^{\text{ES}}$ values averaged over four PFAS compounds from the 3 ps production period.

E.4 Battery Binder

For the third application: Battery Binder. The query as shown in D.1 is:

Battery Binder

```
Which advanced binder technologies are being developed to improve the
↪  performance and longevity of lithium-ion batteries?
```

The best commercial solution is: **ChatGPT_o4_mini_high_DR**.

The best local solution is: **DToR_qwq32B_local500**.

The expert based on the two reports, provided 4 commercial DR report derived candidates (A1-A4), and 4 local DR report derived candidates (B1-B4), visualized in DL_3 1:

“‘latex

NCM811 cathode description

In order to create a test bench for lithium ion battery binder, we modeled a (104)-terminated slab of layered $LiNi_{0.8}Co_{0.1}Mn_{0.1}O_2$ (NCM811) cathode in a periodic cell with a = 27.67 Å, b = 25.55 Å, c = 49.47 Å and $\alpha = \beta = 90°$, $\gamma = 72.51°$ (P1). The atomic slab spans ≈12.1 Å along the surface normal, leaving ≈37.4 Å of vacuum to suppress spurious interactions; the model contains 732 atoms (Li_{174} Ni_{129} Co_{21} Mn_{21} O_{387}). For structural relaxations, we used a large in-plane surface cell (tripled along b) and constrained the bottom ≈1/3 of atoms to mimic bulk termination.

PVDF benchmark binder

For the benchmark binder, we built an all-trans PVDF 10-mer $(–CH_2–CF_2–)_{10}$, end-capped with H/F, pre-relaxed in vacuo, and then laid it flat on the pre-relaxed NCM811 (104) slab, constructed and constrained following our surface-build script. In the configuration, the chain backbone lies approximately parallel to the surface (∼30° off the b-axis), with a closest polymer–surface contact of ∼2.0 Å, an in-plane backbone span of ≈23 Å, and a projected convex-hull footprint of ≈125 Å^2; the polymer fragment contributes $C_{20}H_{21}F_{21}$, bringing the full model to 794 atoms.

A1 — PVPA (after binder adsorbed)

A poly(vinyl-phosphonic acid) 10-mer ($C_{20}H_{32}O_{30}P_{10}$) was adsorbed on the NCM811 (104) slab, using the same surface-build constraints as our base slab. The slab portion is 12.58 Å thick, leaving 36.88 Å of vacuum; the full system contains 824 atoms (slab = 732; polymer = 92). The PVPA backbone lies nearly parallel to the surface (≈34° off the b-axis), spans 20.59 Å in-plane, and covers a projected convex-hull footprint of 115.09 Å^2. Phosphonate O atoms approach surface cations with a shortest O-Li contact of 2 Å, alongside O-O hydrogen-bond contacts to surface oxygen of 2.5–2.9 Å.

A2 — Alg-Ca (after binder adsorbed)

A Ca^{2+}-cross-linked alginate 5-mer (G–M blocks; carboxylates protonated except those coordinating Ca^{2+}) was laid approximately parallel to the NCM811(104) surface with the bottom oxide layers fixed. The combined model has 834 atoms (slab = 732; binder = 102, ≈ $C_{30}H_{40}O_{31}Ca$). The slab portion is 12.20 Å thick with 37.27 Å vacuum. The polymer backbone aligns quasi-parallel to the surface (24.62° off **b**), spans 17.71 Å in-plane, and presents a convex-hull footprint of 92.64 Å^2. Interfacial contacts are dominated by non-specific physisorption (nearest heavy-atom approach 3.23 Å; O_{binder}-O_{slab} 3.31–4.17 Å), while Ca^{2+} remains intrachain-coordinated (closest Ca–O_{slab} 7.13 Å).

A3 — Sulfonyl-polyimide (S-PI) (after binder adsorbed)

A tri-repeat S-PI fragment ($C_{42}H_{14}N_6O_{20}S_3$, 85 atoms) was constructed following the imide-dianhydride/diamino-sulfone motif and combined with the standard NCM811(104) slab (total 817 atoms; slab = 732, binder = 85). The polymer lies parallel to the surface, oriented 37.49° off the b-axis (principal in-plane span 25.71 Å) with a projected convex-hull footprint of 152.10 Å^2. This configuration represents the separated reference: the nearest polymer heavy atom is 17.42 Å above the top oxide layer (no interfacial contacts; shortest $O_{\text{binder}} \cdots O_{\text{slab}}$ = 19.17 Å; $O_{\text{binder}} \cdots$ Ni/Co/Mn = 19.75 Å).

A4 — PTFE $(–CF_2–CF_2–)_{10}$ (after binder adsorbed)

A perfluoroethylene decamer ($C_{20}F_{42}$, 62 atoms) is laid flat on the NCM811(104) surface to provide a non-reactive binder baseline; the combined model has 794 atoms (slab = 732; polymer = 62). The slab portion is 12.20 Å thick with 37.27 Å vacuum. The chain axis lies 39.61° off **b** within the **ab** plane, with a principal in-plane span of 25.14 Å and a projected convex-hull footprint of 80.55 Å^2. Interfacial contacts are dispersive: the nearest heavy-atom approach is 2.49 Å (F-Li); the shortest F-O_slab and F-Ni contacts are 2.84 Å and 3.43 Å, respectively, while C-O_slab $\geq$ 3.63 Å and no metal–fluorine contact falls below 2.49 Å. These metrics set a PTFE benchmark for subsequent ΔE_ads/A comparisons.

B1 — Schiff-base polymer (imine-linked poly(aldehyde-amine)) (after binder adsorbed)

An imine-linked poly(aldehyde-amine) fragment ($\mathbf{C}_{40}\mathbf{H}_{28}\mathbf{N}_4$, 72 atoms) is adsorbed on the NCM811(104) slab; the combined model contains **804** atoms (slab = 732; binder = 72). The oxide portion is **13.07 Å** thick with **36.39 Å** vacuum. The polymer lies quasi-parallel to the surface, oriented **36.97**° off **b** within the **ab** plane; its principal in-plane span is **34.60 Å**, and the projected convex-hull footprint is **168.51 Å**^{2}. Interfacial contacts show a nearest heavy-atom approach of **2.57 Å** (C-Li); the imine N approaches surface oxygen at **2.35 Å**, while separations to cation sites remain larger (N-Li **3.68 Å**; N-Ni **4.29 Å**; N-Co **4.57 Å**; N-Mn **4.76 Å**).

B2 — lignin + CNT (after binder adsorbed)

A β-O-4 lignin dimer was positioned at the NCM811(104) interface and co-laid with an armchair (6,6) carbon nanotube segment (8 axial repeats; 192 C) oriented roughly parallel to the terrace; the tube's in-plane principal axis is rotated **36.86**° off **b** and spans **24.91 Å**. The composite binder portion contains $\mathbf{C}_{234}\mathbf{H}_{80}\mathbf{O}_{46}$. The oxide portion spans **11.05 Å** in thickness along the surface normal with **38.42 Å** of vacuum retained. Lignin's phenolic/ether oxygens face the cation-terminated ridges, with the closest **O-(Ni/Co/Mn)** approach of **3.32 Å**, consistent with physisorptive H-bond/ion-dipole contacts rather than direct chelation; the CNT lies flat to provide an extended π-contact, with a near-surface carbon convex-hull footprint of **84.71 Å**^{2}. Together, this layout tests cooperative anchoring: localized polar adhesion from lignin complemented by broad dispersive contact from the nanotube across the NCM811 terrace.

B3 — Lignin–CMC hybrid (β-O-4 lignin dimer + CMC monomer, Na^+ exchanged) (after binder adsorbed)

The composite contains **876 atoms** (slab = 732; binder = 144 with $\mathbf{C_{50}H_{69}O_{24}Na_1}$) and uses the same NCM811(104) platform and assembly recipe (lignin dimer H-bonded to a CMC unit, carboxylate presented to the terrace). The organic complex lies quasi-parallel to the surface, its principal in-plane axis oriented **38.18°** off **b**, with a **35.09 Å** backbone span and a projected convex-hull footprint of **235.11 Å^2**. The oxide portion is **11.09 Å** thick with **38.38 Å** of vacuum along **c**. Interfacial approach distances indicate mixed ionic/coordination contacts: nearest heavy-atom approach **2.25 Å** (O_binder-Li), with **O_binder-Ni = 2.99 Å**, **O_binder-Co = 4.09 Å**, and **O_binder-Mn = 4.29 Å**; the exchanged **Na^+** sits remote from the terrace (**Na-O_slab = 6.80 Å**). During relaxations, the **bottom one-third of oxide atoms** are held fixed to emulate a semi-infinite substrate.

B4 — ZIF-8 face (Zn_4(2-mim)$_8$ cluster) (after binder adsorbed)

We approximate the MOF by a single sodalite-cage face—four Zn nodes bridged by eight 2-methylimidazolates—so the interface samples the correct local Zn–N_4/μ_2-imidazolate chemistry without imposing the artificial pore matching and long-range lattice constraints of a periodic ZIF on a finite oxide slab. The internal Zn–N network is kept near its intrinsic geometry via mild constraints, while only the interfacial coordinate is relaxed; one Zn is oriented toward a surface oxygen with an initial Zn-O set to ~2.10 Å to seed chemisorption and allow secondary N-Li/C-O_slab contacts to adjust around it. This charge-neutral ring thus captures the short-range bonding physics that governs adhesion at comparable cost to polymer fragments, enabling fair ranking by area-normalized adsorption energy ΔE_ads/A across all candidates.

Battery Binder Adhesion Performance

The mechanical integrity of lithium-ion battery cathodes critically depends on binder adhesion strength, particularly under the chemically aggressive conditions of liquid electrolyte environments. To evaluate candidate binder materials proposed by the DR agents (Figure DL_3 1), we employed the same computational framework described for PFAS sensors, with binding energies normalized by the interfacial contact area A to account for varying polymer–cathode interaction footprints. The contact area was determined through convex hull analysis of atoms within 3.5 Å of the NCM811 surface, enabling direct comparison of adhesion strength per unit interfacial area.

Static DFT calculations reveal favorable adhesion for all candidates, with areal-normalized binding energies ranging from 0.015 to 0.111 eV/Å^2 (Figure DL_3 2). These values, equivalent to 0.24–1.78 J/m^2, align well with established benchmarks for dry interface adhesion, such as graphene on SiO_2 (~0.45 J/m^2) and metal–oxide contacts (~2 J/m^2). Candidates B4 and B1 demonstrate the strongest adhesion, while the benchmark PVDF falls in the mid-range, suggesting potential improvements over current commercial binders.

However, the introduction of explicit battery electrolyte (EC:EMC:$LiPF_6$ = 9:9:2:2 at 1.25 g/cm^3) through AIMD simulations dramatically alters the adhesion landscape (Figures DL_3 3 and DL_3 4). Following the same protocol of 2 ps equilibration and 3 ps production runs, most candidates exhibit negative $\Delta E^{\mathrm{ES}}_{\mathrm{pot}}/A$ values, indicating that the separated, electrolyte-wetted interface becomes energetically favorable over the bound state. This solvent-induced adhesion reversal reflects competitive adsorption of electrolyte molecules at the binder–cathode interface, disrupting direct polymer–surface interactions.

Notably, two exceptions emerge from this trend: candidate A3 (sulfonyl-polyimide) and B2 (lignin with CNT) maintain positive $\Delta E^{\mathrm{ES}}_{\mathrm{pot}}/A$ values even in explicit electrolyte, suggesting robust wet adhesion. This resilience likely arises from multidentate polar anchoring groups and extended π-conjugated contacts that effectively displace the first electrolyte layer. The performance scoring normalized to PVDF baseline (Figures DL_3 5 and DL_3 6) quantitatively captures this wet-interface reordering, where candidates optimized for dry adhesion may fail under operational electrolyte conditions while others demonstrate unexpected electrolyte compatibility. These findings underscore the critical importance of explicit-solvent validation in battery binder design, as interfacial energetics in liquid electrolyte environments cannot be reliably predicted from vacuum or implicit-solvent calculations alone.

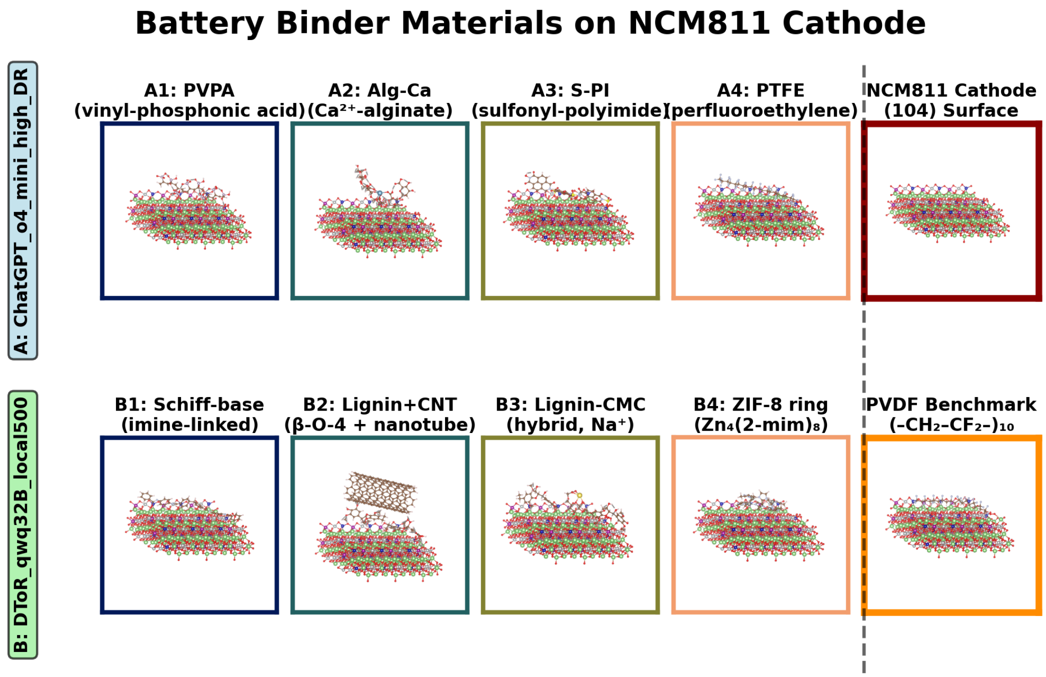


Figure DL_3 1: Structural configurations of battery binder candidates proposed by DR agents, with PVDF [50] included as the domain benchmark reference. Optimized adsorption configurations on the NCM811 cathode surface are shown for each candidate system.

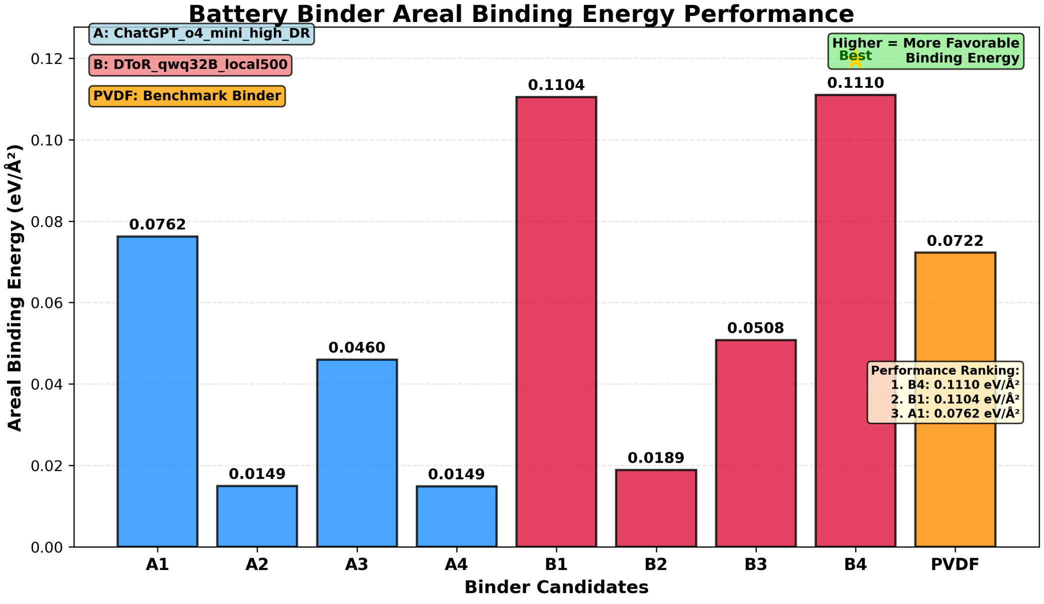


Figure DL_3 2: Comparison of areal-normalized binding energies ($\Delta E_{\text{bind}}/A$) from static DFT calculations. Values are expressed in eV/Å^2, where A represents the interfacial contact area determined by convex hull analysis of atoms within 3.5 Å of the NCM811 surface.

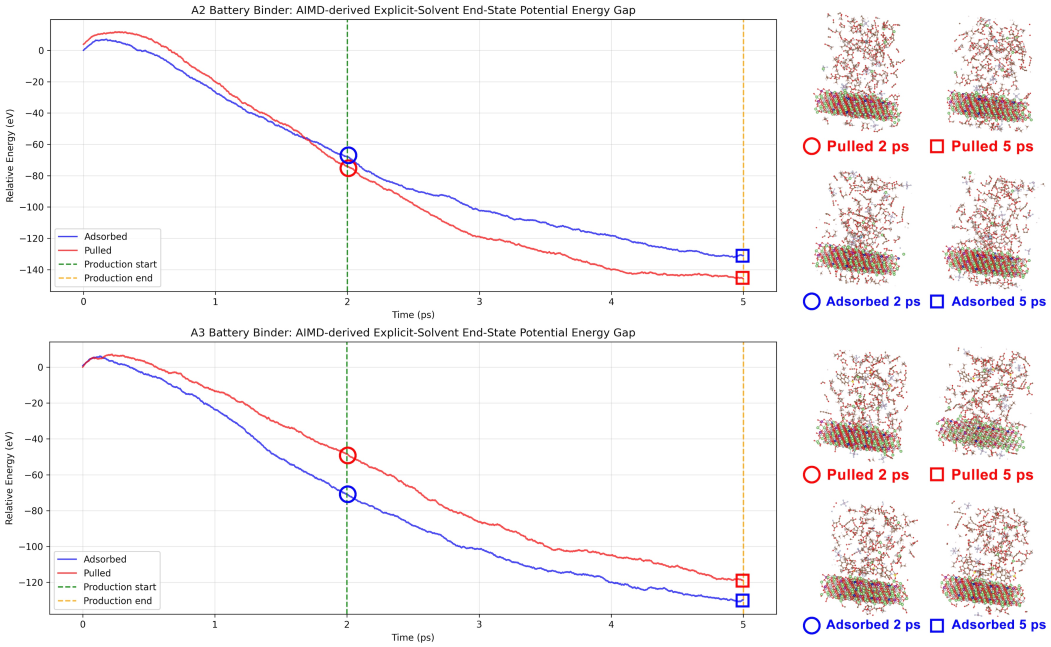


Figure DL_3 3: Representative AIMD trajectory showing system energy evolution in explicit battery electrolyte (EC:EMC:LiPF$_6$ = 9:9:2:2). The green dashed line denotes the 2 ps equilibration period. Right panels display system configurations at highlighted time points along the 3 ps production trajectory.

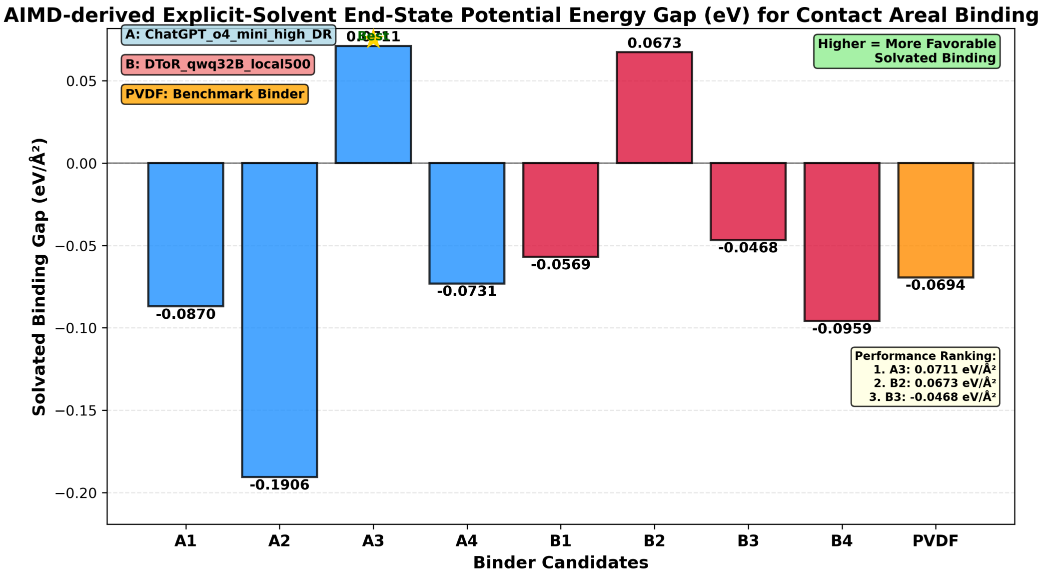


Figure DL_3 4: Areal-normalized explicit-solvent end-state potential energy gaps ($\Delta E_{\mathrm{pot}}^{\mathrm{ES}}/A$) from AIMD simulations. Negative values indicate unfavorable adhesion in the presence of electrolyte (Pulled state in explicit solvation stablizes the pulled state), reflecting competitive solvent adsorption at the binder–cathode interface.

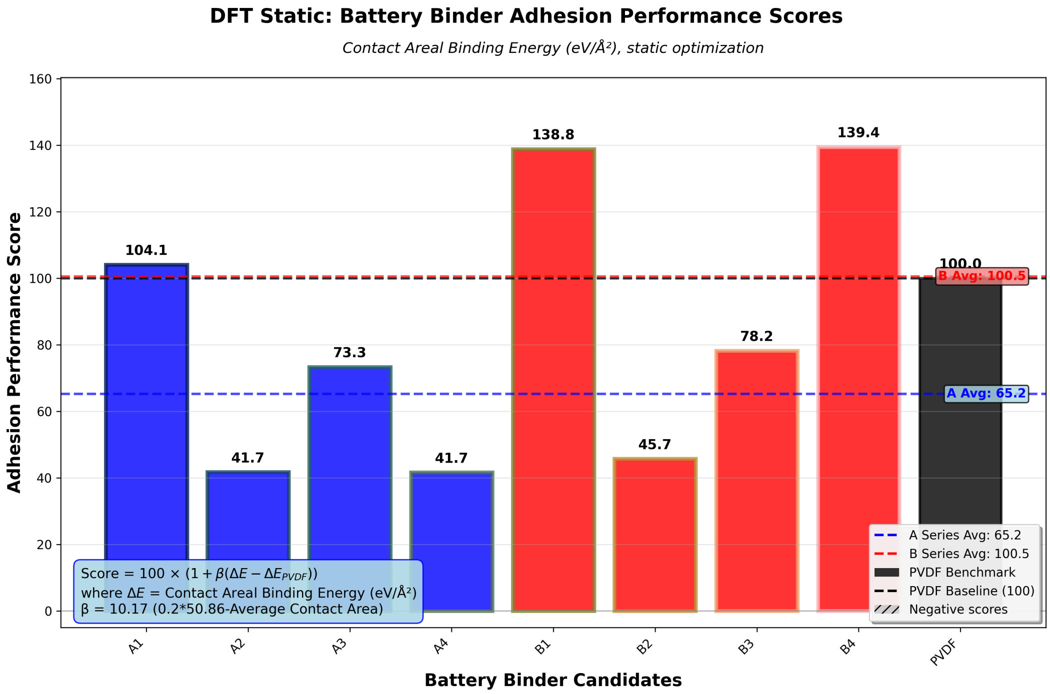


Figure DL_3 5: Adhesion performance scores for all binder candidates from static DFT calculations, normalized to PVDF baseline (100). Scores are calculated using the linear scaling metric with sensitivity parameter $\beta = 10.17$, applied to areal-normalized binding energies.

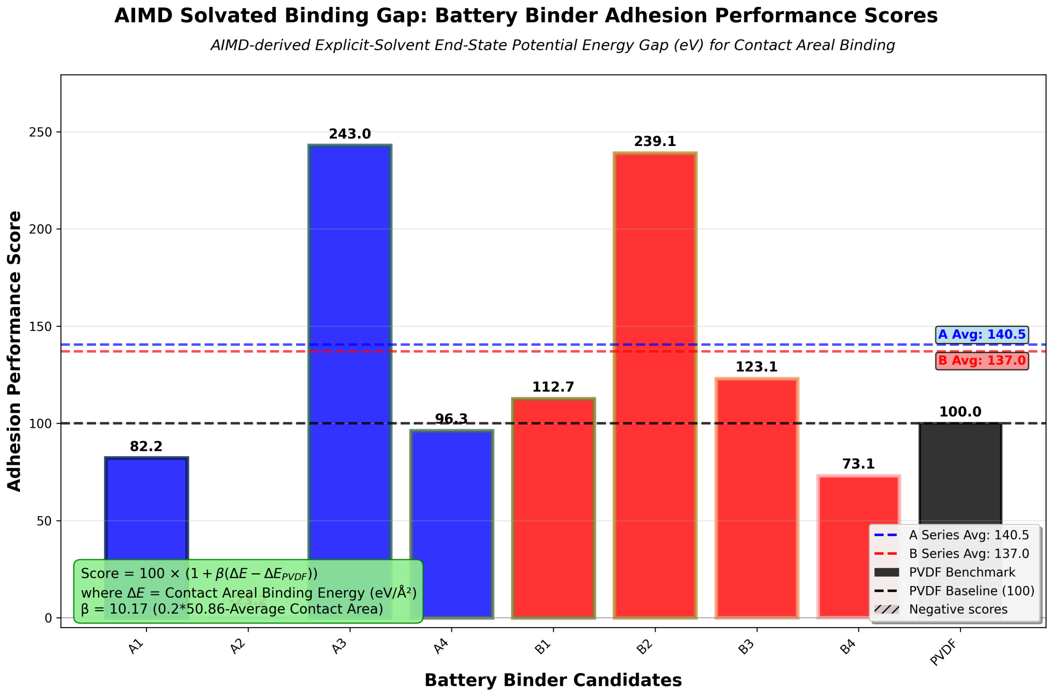


Figure DL_3 6: Adhesion performance scores for all binder candidates from explicit-electrolyte AIMD simulations, normalized to PVDF baseline (100). The scoring methodology remains identical to static DFT, with $\Delta E_{\text{bind}}/A$ replaced by time-averaged $\Delta E^{\text{ES}}_{\text{pot}}/A$ from the 3 ps production period.

E.5 OER Stability

For the fourth application: OER Stability. The query as shown in D.1 is:

OER Stability

```
Identify the top nanostructured electrocatalyst materials for driving the oxygen
↪   evolution reaction (OER) in complex wastewater matrices-such as
↪   high-chloride, high-organic-load, or multi-ion streams-,  for each the
↪   material class/composition, comprehensively consider key performance metrics
↪   (overpotential at 10 mA cm⁻², Faradaic efficiency, stability), and the
↪   surface-engineering strategies that confer corrosion resistance and
↪   sustained activity in real effluent conditions.
```

The best commercial solution is: **Perplexity_DR**.

The best local solution is: **DToR_qwq32B_local500**.

The expert based on the two reports, provided 4 commercial DR report derived candidates (A1-A4), and 4 local DR report derived candidates (B1-B4), visualized in DL_4 1:

Benchmark 110 IrO$_2$

For benchmark material, we created most commonly seen, a stoichiometric rutile IrO$_2$ (110) slab built as a 6 × 3 in-plane supercell along × $[1\bar{1}0]$, with cell parameters **a = 19.06 Å, b = 19.11 Å, c = 45.18 Å,** $\alpha = \beta = \gamma =$ **90° (P1)**, from which the underlying bulk metrics map to $a \approx 4.51$ Å and $c \approx 3.18$ Å. The slab comprises **five O–Ir–O trilayers** (thickness ≈ 15.18 Å) separated by ≈ **30.0 Å** of vacuum along c, yielding **symmetric, O-terminated (bridging-O) surfaces** that expose the conventional row motif used for OER modeling (adjacent Ir$_{\text{cus}}$/O$_{\text{br}}$ sites). The model is **strictly stoichiometric (Ir:O = 1:2; 180 Ir and 360 O)** and thus suitable as a clean, non-polar benchmark surface for computational electrocatalysis studies.

A1 — Ni–Fe layered double hydroxide (LDH) slab (basal)

A hexagonal brucite-derived Ni–Fe hydroxide (001) slab was constructed as a **6** × **6** in-plane supercell (basal lattice $a_0 \approx 3.126$ Å), with cell metrics **a = b = 18.7560 Å, c = 62.3777 Å,** $\alpha = \beta =$ **90°,** $\gamma =$ **120° (P1)**. The model contains **three brucite-like** $[M(OH)\_6]$ octahedral sheets stacked along c (36 cation sites per sheet), with a **Ni:Fe = 81:27** distribution (**25% Fe on the cation sublattice; 9 Fe per sheet**), and overall stoichiometry **Ni_81Fe$_{27}$O$_{216}$H$_{216}$ (540 atoms)**. Both basal surfaces are **symmetrically hydroxyl-terminated** (H atoms constitute the outermost layers), yielding a **stoichiometric, non-polar** slab of thickness ≈ **19.67 Å**, separated by ≈ **42.71 Å** of vacuum within the **c** dimension; the **inter-sheet metal-plane separations** are **7.87** and **7.81 Å**, consistent with LDH-like basal spacing. This geometry provides equivalent top/bottom (001) terminations exposing M–OH motifs suitable for benchmarking OER intermediates on LDH surfaces.

A2 — phosphate-intercalated Ni–Fe LDH (001)

Kept identical to **A1** in lattice and slab construction (6 × 6 hexagonal basal supercell; $a = b =$ 18.756 Å, $\gamma =$ 120°, $c =$ 62.3777 Å; P1) with symmetric hydroxyl-terminated basal surfaces, **A2 is derived from A1 by inserting protonated phosphate molecules (H_3PO_4)** into the interlayer galleries—**18 PO_4 units per supercell (9 per gallery)**, giving an overall composition **Ni_81Fe$_{27}$O$_{288}$H$_{270}$P$_{18}$** (684 atoms, Ni:Fe = 3:1). The intercalation shifts the metal-plane separations from ~7.87/7.81 Å (A1) to **8.89/8.90 Å**, increasing the slab thickness to **22.22 Å** and leaving ~**40.16 Å** of vacuum along c; the phosphorus centroids sit near the mid-planes of each gallery.

A3 — Ni single atoms on Sb-doped SnO_2(110)

A rutile SnO$_2$(110) slab was built as a **4** × **8** in-plane supercell along **[1$\bar{1}$0]** × (cell **a = 26.8022 Å, b = 25.4960 Å, c = 57.0945 Å; α = β = γ = 90°; P1**), comprising **four O–Sn–O trilayers** (slab thickness ≈ **12.72 Å**) and ≈ **44.37 Å** of vacuum. **Eight Sb atoms** substitute Sn **in the subsurface cation plane (second layer from the top)**—**12.5 %** within that plane, **3.2 %** overall (**Sn_244Sb$_8$**). The **top (110) surface** carries **four isolated Ni adatoms** arranged in a 2 × 2 array (**coverage** ≈ **0.585 Ni nm^{-2}**), each adopting a **square-pyramidal NiO_5** environment with **four equatorial Ni–O = 1.922 Å** and **one apical Ni–O = 2.033 Å**; the **bottom surface** is left **clean and O-terminated**. The full supercell stoichiometry is **Sn_244Sb$_8$Ni$_4$O$_{508}$ (764 atoms)**.

A4 — Ce-doped, oxygen-deficient spinel NiCo_2O_4 slab (asymmetric OH termination)

Modeled in **P1** with an oblique in-plane cell (**a = 19.807 Å, b = 30.866 Å,** γ **= 108.715°; c = 41.150 Å**), giving an in-plane area of **579.05 Å^2 (5.79 nm^2)**; the atomic slab spans **14.87 Å** along c and is separated by **26.28 Å** of vacuum. The supercell composition is **Co_188Ce$_8$Ni$_{92}$O$_{376}$H$_{96}$ (760 atoms)**, i.e., **Ni:Co** ≈ **1:2.04** with **Ce at 2.8 at.% of the cation sublattice**. Relative to the ideal AB$_2$O$_4$ stoichiometry for 288 cations, the model contains **8 lattice-oxygen vacancies (x** ≈ **0.021 in NiCo_2O_4** $-\ x$), corresponding—if evenly split between the two faces—to ≈ **0.69 vac nm^{-2}** per surface. **Ce dopants occupy two equivalent subsurface cation planes** (four Ce per plane), located ∼**4.31 Å below the top** and ∼**5.32 Å above the bottom** surface, preserving an overall centrosymmetric placement. The **top surface is Co-rich and partially hydroxylated** (within the outermost 2 Å: **8 Co, 20 O, 28 H** → ≈ **1.38 Co nm^{-2}** and ≈ **5.5 OH nm^{-2}**), while the **bottom surface is more strongly hydroxyl-passivated** (**64 H** on that side overall, ≈ **11.1 OH nm^{-2}**). No Ni appears within the outermost 2 Å of the top surface (Ni resides in subsurface layers), yielding a **Co–O(H) termination** appropriate for probing OER pathways on spinel-type oxides.

B1 — graphene-capped Ni–Fe LDH (001)

Retaining the **A1** lattice and slab construction (6 × 6 hexagonal basal supercell; a = b = 18.756 Å, γ = 120°, c = 70.6707 Å; P1), this model is derived from **A1** by placing a **single-layer graphene overlayer (128 C)** above the top basal surface while keeping the LDH stoichiometry **Ni_81Fe$_{27}$O$_{216}$H$_{216}$** unchanged. The LDH comprises **three brucite-like [M(OH)_6] sheets** (Ni:Fe = 3:1) with **metal-plane separations of 7.87 and 7.81 Å**; both basal faces are hydroxyl-terminated, but only the **top face** is covered by graphene. The LDH portion spans **19.67 Å** along c; including graphene, the occupied thickness is **24.67 Å**, leaving **46.00 Å** of vacuum. The graphene sheet is essentially planar (out-of-plane corrugation ≤ 0.001 Å) and sits **5.00 Å above the outermost hydroxyl-H layer** (**5.97 Å above the topmost O layer**), consistent with a physisorbed, non-perturbative interface that preserves the LDH surface structure; the **bottom layers are identical to A1** and remain exposed/hydroxylated for benchmarking.

B2 — graphene-capped hematite (α-Fe_2O_3) (0001) with an aperture

A hexagonal α-Fe$_2$O$_3$(0001) slab was modeled in **P1** with **a = b = 20.152 Å,** γ **= 120°; c = 60.339 Å** (in-plane area **351.70 Å^2 = 3.517 nm^2**), giving an atomistic thickness of **32.20 Å** separated by **28.14 Å** of vacuum. The slab is **stoichiometric Fe_2O_3** in the interior (**Fe_384O$_{576}$**) with **both faces hydroxyl-terminated (H_92 total)** to render a non-polar model. A **single sp^2-carbon overlayer (C_115)** caps only the **top** face and includes a **circular** ∼**6 Å "window"** (aperture) that leaves a patch of **exposed surface Fe sites** beneath the opening; elsewhere the carbon sheet remains graphene-like, with its centroid ∼**2.84 Å above the outermost O plane** (≈ **2.04 Å above terminal H**) and a small out-of-plane corrugation (**Δz** ≈ **0.91 Å**). The **bottom** face remains hydroxylated and exposed for benchmarking.

B3 — BaCo_0.6Fe$_{0.4}$O$_3$ perovskite (001), BO$_2$||BaO asymmetric slab

In line with Branch B's chloride-resilience goal, a stoichiometric Ba(B)O$_3$ slab with mixed B-site cations (Ba$_{180}$Co$_{108}$Fe$_{72}$O$_{540}$; 900 atoms) was modeled in P1 with **a = b = 23.820 Å, c = 49.850 Å, $\alpha = \beta = \gamma$ = 90°** (in-plane area **567.39 Å^2 = 5.674 nm^2**). The atomic slab spans **19.31 Å** along *c* and is separated by **30.54 Å** of vacuum. Terminations are **asymmetric**: the **top** face is a **BO_2 layer** that is **Co-rich (Co:Fe = 27:9 $\approx$ 3:1)**, while the **bottom** face is **BaO-terminated**; within ±2 Å of the surfaces this corresponds to $\sim$**6.35 B nm^{-2}** (with $\sim$**12.69 O nm^{-2}**) at the top and $\sim$**6.35 Ba nm^{-2}** at the bottom. This construction furnishes a corrosion-tolerant, **bare BO_2 active surface** over a BaO support, consistent with the initial design brief to benchmark robust oxide surfaces against chloride attack.

B4 — coral-mimetic Ni–Fe LDH (001)

Keeping the MXene/LDH base identical to **A1/B1** (6 × 6 hex basal supercell; *a* = *b* = 18.756 Å, γ = 120°, *c* = 62.3777 Å; P1), the "coral" morphology was introduced by excising a single cylindrical pore ($\sim$8 Å diameter) through the **top two** hydroxide layers and hydroxyl-passivating all newly exposed edges, as specified in the design note; both basal faces remain hydroxyl-terminated, with the **bottom face unchanged** from A1/B1. The resulting model is **Ni_69Fe$_{25}$O$_{192}$H$_{216}$ (502 atoms)**, occupying **26.58 Å** along *c* with **35.79 Å** vacuum (in-plane area **304.66 Å^2**), yielding a non-polar, porous benchmark that preserves the original basal registry while exposing M–OH rim sites within the aperture.

Metal Site Choices Considering the complexity of each slab created based on DR report, we decide to select from each slab only representative sites for further study:

IrO_2

On IrO$_2$(110) we target the canonical fivefold Ir terrace site (Ir$_{5c}$) on the top surface cation row. Its local geometry is square-pyramidal, with four coplanar surface O neighbors at 2.00 Å and a single subsurface apical O at 1.97 Å; the missing apical ligand points toward vacuum. Situated in the outermost O–Ir–O trilayer, this terrace Ir is the standard benchmark adsorption site on rutile-type oxides and is chosen to probe intrinsic, defect-free reactivity under well-defined coordination.

A1

On A1 we probe two "regular" terrace cations on the **top basal (001) surface**—one Fe and one Ni located in the central, step-free region. Both exhibit near-ideal octahedral **M(OH)_6** coordination with a 3-up/3-down O arrangement; their mean M–O distances are **2.09 Å (Fe)** and **2.07 Å (Ni)**, and the in-plane metal–metal spacing reflects the basal lattice at $\sim$**3.13 Å**. The Fe site sits in a Ni-rich cation shell (isolated Fe amid Ni), whereas the Ni site has a typical mixed neighborhood ($\sim$4 Ni, 2 Fe), consistent with the $\sim$3:1 Ni:Fe layer composition. These terrace positions were chosen as **representative, bulk-like references**—well coordinated and far from edges or defects—providing robust baselines for adsorption benchmarking on LDH surfaces.

A2

In **A2** we focus on two regular first-layer cations on the **top basal (001) surface**—one Fe and one Ni—above the phosphate-filled interlayer. Both sites are octahedral with **3 O above / 3 O below**; the Fe exhibits **Fe–O** ≈ **2.06/2.01 Å** (above/below) and the Ni **Ni–O** ≈ **2.10/2.07 Å**. Relative to the subsurface phosphate, the Fe sits **nearly on-axis** over a PO_4 unit (**Δz** ≈ **4.30 Å**, lateral offset ≈ **0.71 Å**), whereas the Ni is **laterally off-axis** from the nearest PO_4 (**Δz** ≈ **4.54 Å**, lateral ≈ **2.95 Å**). These two surface metals are chosen as representative, symmetry-inequivalent terrace sites that sample **distinct interlayer electrostatic environments** (aligned vs off-axis to PO_4) while avoiding edges/defects—providing a solid, structure-based rationale for benchmarking metal-identity and phosphate-proximity effects.

A3

For A3, on the top SnO_2(110) terrace, the representative site comprises a single Ni atom anchored above a bridging-O row, adopting a square-pyramidal NiO_5 geometry with four surface O_{br} ligands in the equatorial plane (Ni–O ≈ 1.92 Å) and one apical O beneath (Ni–O ≈ 2.03 Å); the Ni sits ∼0.19 Å below the O_{br} plane and ∼2.03 Å above the apical O. The two nearest subsurface Sn cations on adjacent cation rows are each connected to the adatom via a Sn–O_{br}–Ni bridge (Sn–O ≈ 2.07 Å; Ni–Sn ≈ 3.12 Å). Situated on the upper, flat terrace away from step edges, this motif matches your original "Ni single-atom adjacent to Sn via bridging O" note and is selected as a representative active site because it captures the canonical O_{br}–M_{cus} environment of rutile(110) while providing quantitatively well-defined, low-coverage coordination ideal for benchmarking OER intermediates.

A4

For A4, we have choosed a total of 6 Co/Ni atoms. On the top Co–O(H)-terminated face of the Ce-doped, oxygen-deficient $NiCo_2O_4$ slab, we consider two Ce-proximal motifs and one Ce-remote control set. Ce-proximal: (i) a two-fold-coordinated Co sitting at the rim of a surface O-vacancy directly above the subsurface Ce layer, and (ii) a three-fold Co one O row back on the same terrace; alongside these, a three-fold Ni in the same Co-rich layer but still within the Ce-perturbed field captures the B-site identity contrast. Ce-remote (terrace center, far from dopants/vacancies): (iii) a two-/three-fold Co pair representative of the intact hydroxylated terrace, and (iv) a two-fold Ni surrounded by a largely unreconstructed O sublattice. Together these five local environments span coordination (2 vs. 3), cation identity (Co vs. Ni), and dopant/vacancy proximity (near-Ce vs. terrace-center) consistent with the A4 design (Ce in subsurface planes + engineered surface O-vacancies with hydroxyl passivation). This refelects the selection logic: Ce-adjacent rims/near-rim terrace sites plus Ce-remote terrace controls with 2- and 3-coordinate Co/Ni.

B1

On **B1 (graphene-capped NiFe-LDH (001))**, the representative **first-layer surface Fe and Ni** are the outermost cations in the top brucite-like sheet directly beneath the graphene overlayer. Each sits in a **regular M(OH)_6 octahedron** with nearly isotropic M–O bonds (Fe: **2.09 Å** apical/**2.09 Å** apical; equatorial **2.09** ± **0.00 Å**; Ni: **2.08 Å** apical/**2.08 Å** apical; equatorial **2.07** ± **0.00 Å**; octahedral distortion ≈ **0.11 %**), positioned **1.03 Å** below the topmost O plane and **2.00 Å** below terminal OH, with the graphene sheet **7.00 Å** above the metal plane (**5.97 Å** above the surface O layer). The chosen Fe lies near the in-plane center (> **8.09 Å** from periodic edges) and the Ni is also well inside the terrace (≈ **5.45 Å** from edges), i.e., **regular terrace sites** rather than step/edge motifs—matching the "regular first-layer" description in the earlier manual note. These sites were selected to provide **clean, defensible benchmarks** for OER under Branch-B1's design goal (probing LDH activity beneath a protective graphene barrier): they are far from defects and intercalants, exhibit bulk-like octahedral geometry, and experience only **physisorbed** graphene influence, thus isolating intrinsic Fe/Ni–OH reactivity at the **top basal surface**.

B2

on B2, On the top (0001) face beneath the graphene aperture, four representative surface Fe sites were selected: two **rim** Fe located along the edge of the ∼6 Å window where the sp^2 rim approaches the oxide, each making Fe–C contact (shortest Fe–C ≈ 2.17 Å), with under-coordination (Fe–O CN ≈ 3) and a slight outward relaxation (≈ 1.33 Å above the terminal O plane); and two **center-of-window** Fe positioned near the middle of the opening, far from the carbon edge (nearest C $\geq$ 5.24 Å), retaining octahedral-like coordination (Fe–O CN ≈ 6) and sitting ≈ 1.19 Å below the top O/H plane. These two site families—carbon-contacted rim vs carbon-free aperture interior—follow the original design intent and provide a physically motivated contrast to quantify how the graphene shell perturbs intrinsic hematite reactivity.

B3

On the BO$_2$-terminated top surface of B3 we sample four terrace cations: (i) one **Fe** and (ii) one **Co** sitting at the rim of a deliberate surface O-vacancy, each **three-coordinated** to lattice O (vacancy-edge sites, no apical O) with shortened in-plane M–O distances of ∼**1.90–1.97 Å**; and, as references, (iii) one **Fe** and (iv) one **Co** on the same terrace but **far from any defect** ($\geq$**10.00 Å** lateral separation), retaining the regular **square-pyramidal BO_**5 environment (four equatorial M–O ≈ **1.95–2.02 Å**, one apical ≈ **2.05–2.10 Å**). This choice—consistent with the original notes—targets the most contrastive local motifs (vacancy-edge undercoordination vs pristine terrace; Fe vs Co identity) while keeping all sites in the **outermost BO_**2 **plane** above the BaO underlayer.

B4

In B4's coral-type aperture, we target four representative rim sites: a Fe and a Ni in the **top metal plane** at the pore edge, and the corresponding **second-plane** Fe and Ni located directly beneath the rim (as in the earlier manual note). The two top-layer cations sit **1.88 Å (Fe)** and **1.55 Å (Ni)** below the outermost O plane and lie within **0.10–0.23 Å** of the top-plane centroid; their in-plane separation is **3.21 Å**. At the rim, **Fe remains near-octahedral** (**6 O $\leq$ 2.30 Å; $\langle d_{\text{M–O}} \rangle$ = 2.13 Å**), whereas **Ni is under-coordinated** (**4 O $\leq$ 2.30 Å; $\langle d_{\text{M–O}} \rangle$ = 2.12 Å**). The second-layer pair occupies the ring immediately below the aperture edge, **13.53–13.43 Å** beneath the top O plane and within **0.09–0.02 Å** of the second-plane centroid; both are **under-coordinated** (**4 O $\leq$ 2.30 Å; $\langle d_{\text{M–O}} \rangle$ = 2.12 Å (Ni), 2.06 Å (Fe)**), with an in-plane separation of **3.21 Å**. These sites were chosen to systematically span **cation identity (Fe vs Ni)**, **depth (surface vs subsurface)**, and **coordination state at the pore edge**, i.e., the structural factors most likely to govern *OH/*O/*OOH binding and chloride tolerance in the coral design.

OER Catalyst Stability Assessment

Discussion of OER stability. While recent advances in transition metal layered double hydroxides (LDH) and related materials have achieved excellent oxygen evolution reaction (OER) activity approaching theoretical limits, catalyst stability remains the critical bottleneck preventing commercial deployment of water splitting technologies. Operating potentials exceeding 1.5 V vs. RHE in alkaline conditions create thermodynamically favorable pathways for both metal dissolution and lattice oxygen loss, leading to rapid catalyst degradation. To address this fundamental challenge, we evaluated candidate catalysts proposed by the DR agents (Figure DL_4 1) using two complementary stability descriptors following established protocols from our previous work [51]: dissolution potential (U_{diss}) and oxygen vacancy formation energy (V_O).

Stability Descriptors and Physical Interpretation

The dissolution potential U_{diss} quantifies the thermodynamic driving force for metal atom leaching from the catalyst surface, calculated as:

$$U_{\text{diss}}(M@\text{slab}) = U^{\circ}_{\text{diss,bulk}}(M^{z+}/M) + \frac{E_{\text{slab-vac}}(M) + \mu_{\text{bulk}}(M) - E_{\text{slab}}}{ne} \tag{6}$$

where E_{slab} and $E_{\text{slab-vac}}(M)$ represent the DFT energies of the intact and metal-vacancy surfaces, respectively (Figure DL_4 2), $\mu_{\text{bulk}}(M)$ is the bulk metal chemical potential, $U^{\circ}_{\text{diss,bulk}}$ is the standard reduction potential, and n is the number of electrons transferred. Higher U_{diss} values indicate greater resistance to metal dissolution, with positive values suggesting thermodynamic stability under standard conditions.

The oxygen vacancy formation energy V_O captures the structural integrity of the oxide lattice during OER cycling. We following the following protocol for picking lattice oxygen atom to remove, oxygen atoms are identified within 3.0 Å of active metal sites and removed based on coordination number, with under-coordinated oxygen atoms preferentially selected (Figure DL_4 2). Higher V_O values indicate stronger metal–oxygen bonds and reduced susceptibility to lattice oxygen participation in the OER mechanism, which can lead to irreversible structural degradation.

Comparative Stability Analysis

Evaluation across representative surface sites (Figure DL_4 3) reveals significant variation in stability metrics among candidate catalysts. The dissolution potentials span from approximately -0.4 V to $+1.2$ V vs. SHE (Figure DL_4 4). Oxygen vacancy formation energies similarly exhibit broad distribution. In general, higher Vo suggesting enhanced lattice stability potentially arising from stronger metal–oxygen covalency or favorable coordination environments that stabilize the oxide framework.

Correlation Between Stability Metrics

A positive correlation emerges between U_{diss} and V_O across all 27 evaluated metal sites (Figure DL_4 6), indicating that resistance to metal dissolution and oxygen vacancy formation are intrinsically linked. This correlation suggests a unified electronic origin for both stability mechanisms: catalysts with stronger metal–oxygen bonds resist both metal leaching and oxygen loss. The observed relationship validates the use of either metric as a reasonable proxy for overall stability, though evaluating both provides more comprehensive assessment.

Normalized Stability Scoring

To enable quantitative comparison, we similarly applied the linear scoring framework with IrO_2 as the baseline reference (Score = 100), using sensitivity parameter $\beta = 0.2$ consistent with other applications. The U_{diss}-based scores (Figure DL_4 7). The V_O-based scores (Figure DL_4 8) show similar trends but not quite exactly the same rankings, highlighting the value of assessing both metrics.

The divergence between U_{diss} and V_O rankings for specific candidates reflects different stability-limiting mechanisms: some materials may excel at preventing metal dissolution while remaining vulnerable to oxygen loss, or vice versa. This complementary information guides rational catalyst design—for instance, candidates with high U_{diss} but moderate V_O might benefit from strategies to strengthen metal–oxygen bonds, while those with the opposite profile could require enhanced metal site anchoring.

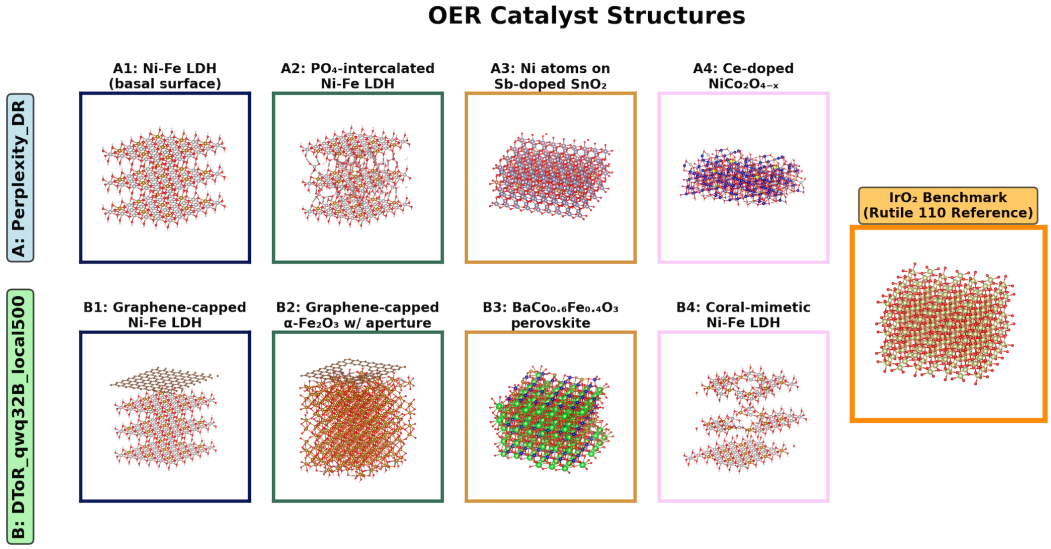


Figure DL_4 1: Structural configurations of OER catalyst candidates proposed by DR agents, with IrO_2 (110) included as the domain benchmark reference for stability assessment.

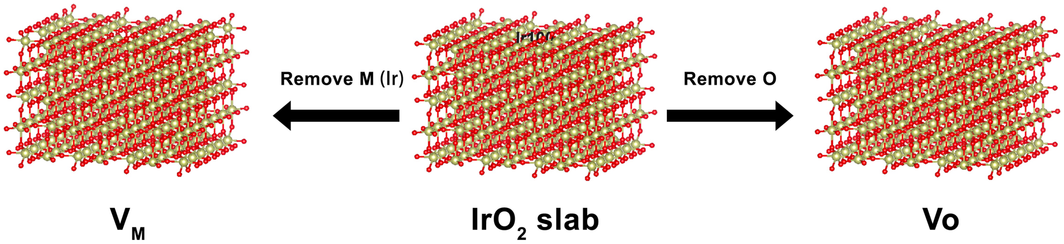
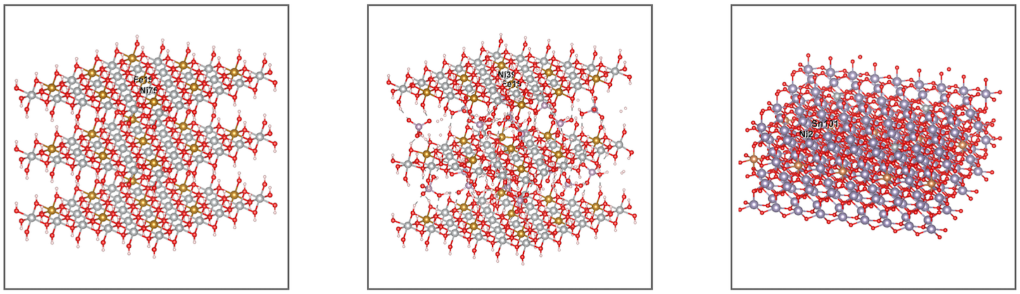


Figure DL_4 2: Schematic illustration of vacancy creation on the IrO_2 rutile (110) surface. Left: Metal vacancy (V_M) created by removing a surface metal atom for dissolution potential (U_{diss}) calculation. Right: Oxygen vacancy (V_O) created by removing a surface oxygen atom for vacancy formation energy calculation.

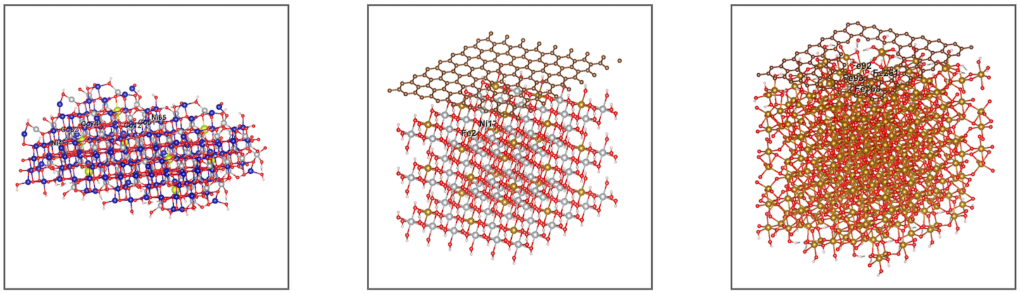
OER Catalyst Site Choices



A1
Regular terrace Fe/Ni on basal (001) LDH with octahedral M(OH)₆ coordination

A2
Fe/Ni above phosphate interlayer with distinct PO₄ proximity effects

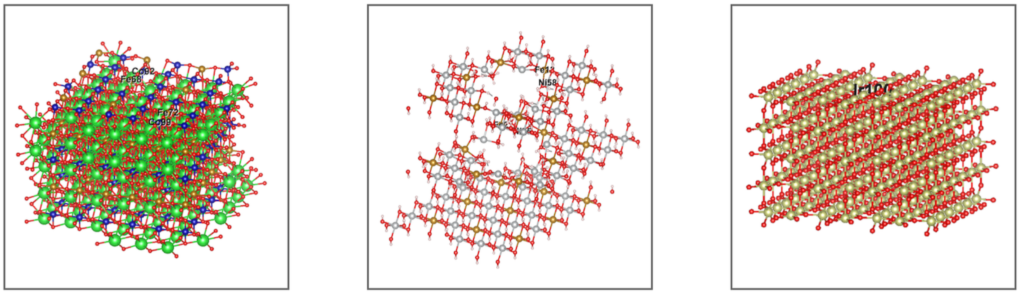
A3
Single Ni atom on SnO₂(110) terrace with square-pyramidal NiO₅ geometry



A4
6 Co/Ni sites: Ce-proximal vacancy rim and terrace sites vs Ce-remote controls

B1
Regular first-layer Fe/Ni beneath protective graphene overlayer

B2
Fe rim sites at graphene aperture vs center-window positions



B3
Fe/Co at O-vacancy edges vs pristine terrace reference sites

B4
Fe/Ni at coral pore rim: top-layer vs subsurface positions

IrO2
Canonical fivefold Ir terrace site on rutile IrO₂(110) benchmark

Figure DL_4 3: Matrix representation of selected metal sites for stability assessment across A-series (A1–A4) and B-series (B1–B4) catalyst candidates. Representative surface metal atoms are highlighted for dissolution and vacancy formation calculations.

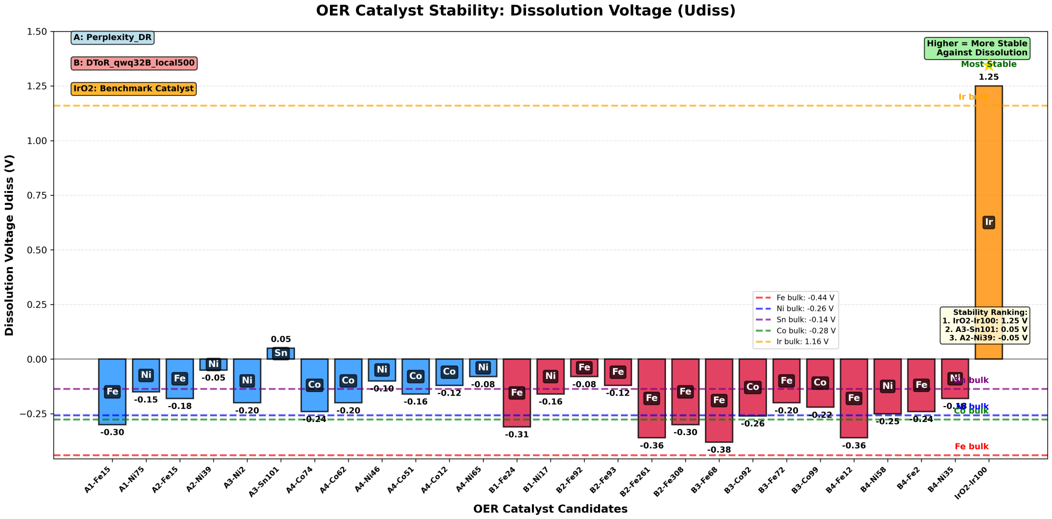


Figure DL_4 4: Dissolution potentials (U_{diss}) for all candidate catalysts at representative surface sites. Higher values indicate greater resistance to metal dissolution. Dashed horizontal lines denote standard bulk dissolution potentials for reference metals (Fe: -0.44 V, Co: -0.277 V, Ni: -0.257 V, Ir: $+1.16$ V vs. SHE).

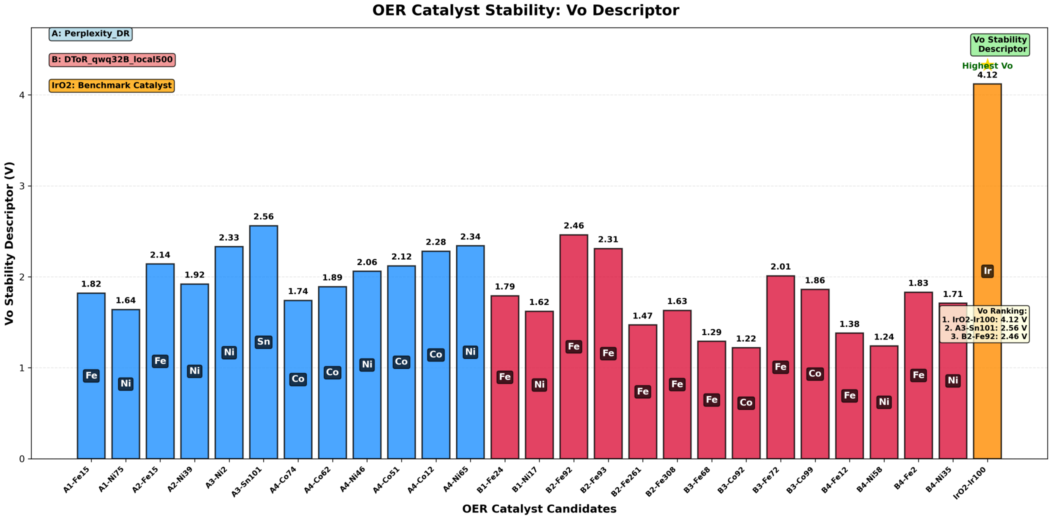


Figure DL_4 5: Oxygen vacancy formation energies (V_O) for all candidate catalysts at representative surface sites. Higher formation energies indicate greater structural stability and resistance to oxygen loss during OER operation.

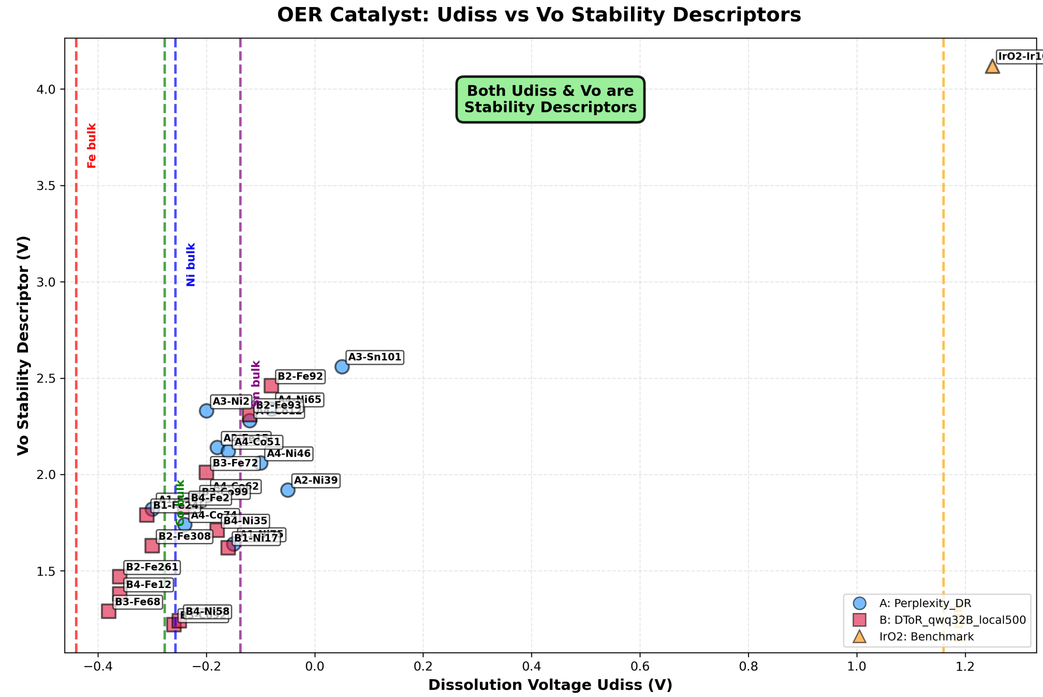


Figure DL_4 6: Correlation between dissolution potential (U_{diss}) and oxygen vacancy formation energy (V_{O}) across all 26 evaluated metal sites. The positive correlation indicates that catalysts resistant to metal dissolution also tend to resist oxygen vacancy formation.

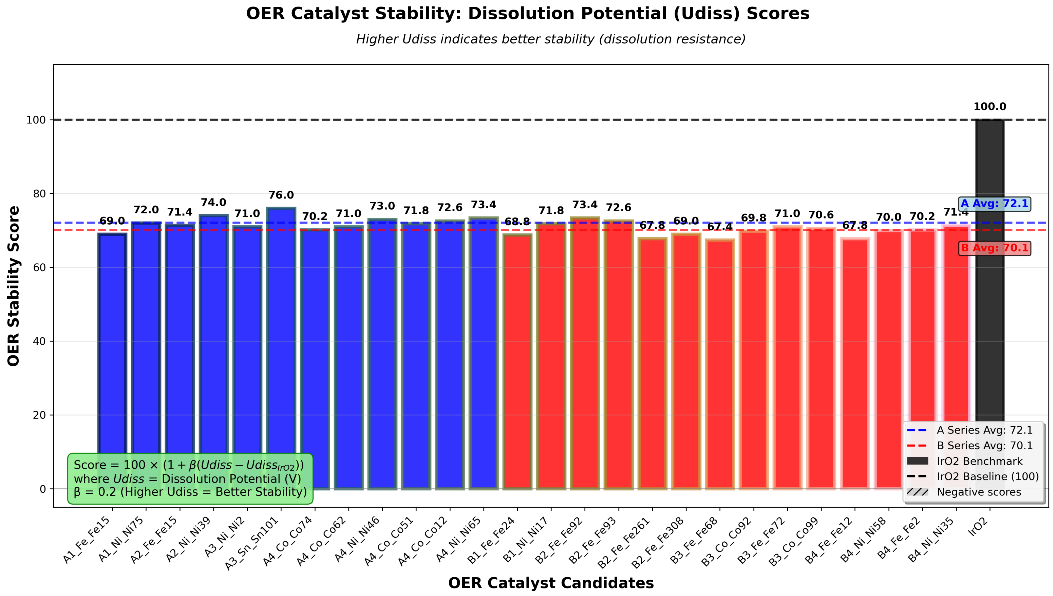


Figure DL_4 7: Stability scores based on dissolution potential (U_{diss}), normalized to IrO_2 baseline (100). Scores are calculated using the linear metric with sensitivity parameter $\beta = 0.2$, where higher U_{diss} values yield higher scores.

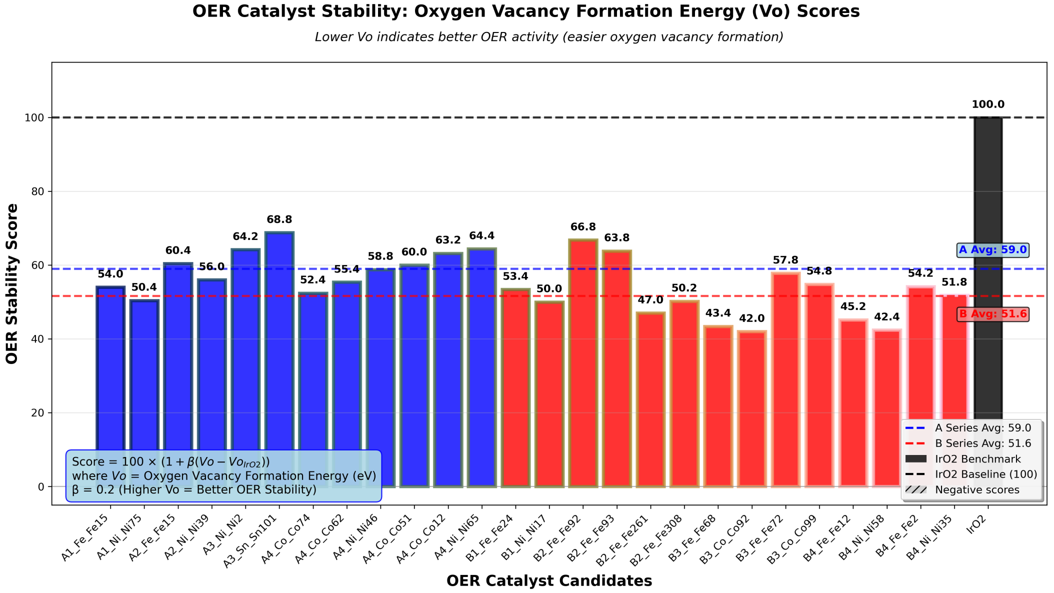


Figure DL_4 8: Stability scores based on oxygen vacancy formation energy (V_O), normalized to IrO_2 baseline (100). The scoring methodology employs the same linear scaling ($\beta = 0.2$), with higher V_O values indicating superior stability.

E.6 CO_2 Sensor Probe

For the fifth application: CO_2 Sensor Probe. The query as shown in D.1 is:

CO2 Sensor Probe

```
Which two-dimensional materials-such as graphene derivatives, transition metal
↪  dichalcogenides, or MXenes-offer the highest CO2 sensing performance in
↪  complex gas or aqueous environments, and how do they compare in terms of
↪  detection limit (ppm), selectivity against common interferents (e.g., O2,
↪  H2O), response/recovery time, and long-term stability, including any
↪  functionalization or structural modifications that enhance these metrics?
↪  Think of novel candidates.
```

The best commercial solution is: **GlideLogic_DR**.

The best local solution is: **DToR_qwq32B_local500**.

The expert based on the two reports, provided 4 commercial DR report derived candidates (A1-A4), and 4 local DR report derived candidates (B1-B4), visualized in DL_5 1:

g-C_3N_4 description

For CO_2 binding, the benchmark substrate is a monolayer graphitic carbon nitride (g-C3N4) built as a 7×7 in-plane supercell of the primitive sheet (49 formula units; C147N196) in a hexagonal cell with a = b = 33.50 Å, γ = 120°, and c = 35.0 Å. All atoms lie in a single atomic plane at z $\approx$ 5.0 Å, leaving ~30 Å of vacuum to the next periodic image along c, i.e., no interlayer interactions. The in-plane network is a fully conjugated sp^2 C–N lattice: each C is three-coordinate, N sites are predominantly two- to three-coordinate ($\langle CN \rangle \approx 2.25$), and nearest-neighbor C–N distances cluster at 1.37 ± 0.06 Å. No heteroatoms, defects, terminations, or adsorbates are present (P1 by construction). This clean, stoichiometric, and strictly planar C_3N_4 slab (areal density $\approx$ 5.05 formula units nm^{-2}) serves as our reference surface for CO_2 binding.

A1 (rGO–PEI)

Hexagonal P1 slab with a = b = 36.90 Å, γ = 120.00°, c = 45.00 Å ($\alpha = \beta$ = 90.00°). The model contains 605 atoms (C461H80O54N10). Atomic z positions span 4.63–16.63 Å, giving a slab thickness of 12.01 Å and 32.99 Å of vacuum along c. The in-plane area is 11.79 nm^2, yielding areal number densities of C 39.09 nm^{-2}, H 6.78 nm^{-2}, O 4.58 nm^{-2}, and N 0.85 nm^{-2}. Carbon heights resolve a dominant rGO basal plane centered at z $\approx$ 7.39 ± 0.64 Å (441 C) and a minor upper carbon domain at 12.09–15.35 Å (20 C). All N atoms reside in the upper domain (11.23–16.53 Å; $\langle z \rangle$ = 13.03 Å), consistent with an amine-rich polyethylenimine (PEI) branch moiety tethered above the rGO sheet; O atoms are concentrated below 10.65 Å, consistent with residual rGO oxygen functionalities. This rGO–PEI slab, isolated by ample vacuum, provides a structurally well-defined platform for site-specific CO_2 binding calculations.

A2 (phosphorene | Al_2O_3 cap with drilled through-pore)

Orthorhombic P1 cell (a = 26.36 Å, b = 27.26 Å, c = 27.08 Å; $\alpha = \beta = \gamma$ = 90.00°) comprising a phosphorene sublayer (P192) beneath an alumina cap (Al146O$_{220}$). Atomic z spans 3.95–17.44 Å, giving a slab thickness of 13.49 Å and 13.59 Å of vacuum. Phosphorene occupies z = 3.95–8.12 Å (thickness 4.17 Å; top P plane at 8.12 Å), while the cap exhibits O/Al = 1.51 ($\approx Al_2O_3$) with the Al framework at z = 8.83–15.83 Å (7.00 Å) and the O network extending 8.05–17.44 Å (9.39 Å), slightly overlapping the top P plane (Δz = 0.07 Å) to form an interfacial contact. The in-plane area is 718.57 $Å^2$ (7.19 nm^2). A single through-pore pierces the cap near (f_x, f_y) = (0.51, 0.50): the top aperture has radius 6.27 Å ($\emptyset$ 12.54 Å), narrows to a bottleneck radius 5.11 Å ($\emptyset$ 10.22 Å; section 82.02 $Å^2$, 11.42

A3 (MoS$_2$ monolayer with vacancies)

P1 cell with a = 27.58 Å, b = 22.29 Å, c = 28.13 Å (α = β = γ = 90.00°). The slab contains 200 atoms (Mo64S136; S/Mo = 2.13) and occupies z = 4.83–8.50 Å (thickness 3.67 Å) with 24.46 Å of vacuum along c. The in-plane area is 6.15 nm^2, giving areal number densities of Mo 10.41 nm^{-2} and S 22.13 nm^{-2}. Layering is a single S–Mo–S tri-layer: the Mo plane is centered at z = 6.56 Å (spread 0.31 Å), bracketed by S planes at z = 5.01 Å and 8.12 Å, yielding an average S–S separation of 3.11 Å and Mo–S vertical offsets of 1.54 Å (bottom) and 1.57 Å (top). No heteroatoms, functional groups, or adsorbates are present. Five Mo atoms are removed to create Mo vacancies. This monolayer MoS$_2$ slab provides a simple, well-isolated reference for CO$_2$ adsorption benchmarks.

A4 (Ti$_3$C$_2$T$_x$, mixed terminations)

Hexagonal P1 slab with a = b = 30.67 Å, γ = 120.00°, c = 51.70 Å (α = β = 90.00°); in-plane area 8.15 nm^2. The model contains Ti300C200O175F25H$_{25}$, i.e., 100 Ti$_3$C$_2$ formula units with 200 surface sites. Terminations are asymmetric: the top face comprises 50.00

B1 (rGO–NH$_2$; same rGO basal slab as A1)

This model reuses the rGO lattice from A1 without change (a = b = 36.90 Å, γ = 120.00°, c = 45.00 Å; in-plane area 11.79 nm^2), retaining the 441-C rippled basal plane centered at z = 7.29 ± 0.67 Å. The slab contains 585 atoms (C441H70O54N20) with heights spanning 4.36–11.16 Å (thickness 6.80 Å) and 38.20 Å of vacuum. Oxygen functionalities lie over 4.36–10.64 Å, while primary amines are grafted exclusively on the upper face (N 8.37–10.38 Å, $\langle z \rangle$ = 9.40 ± 0.63 Å; H up to 11.16 Å), giving an –NH$_2$ areal density of 1.70 nm^{-2}. Areal number densities are C 37.40 nm^{-2}, O 4.58 nm^{-2}, N 1.70 nm^{-2}, and H 5.94 nm^{-2}. With lattice vectors and basal carbon topology identical to A1, this slab isolates the effect of replacing the PEI overlayer with discrete –NH$_2$ terminations for CO$_2$ binding studies.

B2 (Ti$_3$C$_2$F$_2$, fully F-terminated; same construction protocol as A4)

Hexagonal P1 slab with a = b = 30.67 Å, γ = 120.00°, c = 51.70 Å (α = β = 90.00°); in-plane area 8.15 nm^2. The CIF contains Ti300C200F$_{200}$ (100 Ti$_3$C$_2$ formula units; 200 surface sites), i.e., both faces are 100.00

B3 (graphene | Ti3C2O2 heterojunction; same MXene cell as As A4/B2)

Hexagonal P1 slab with a = b = 30.67 Å, γ = 120.00°, c = 51.70 Å (in-plane area 8.15 nm^2). The MXene sub-slab matches the A4/B2 Ti3C2 lattice and is fully *O-terminated on both faces (Ti300C200O200; 100 *O per face, 12.28 sites nm^{-2}). A single-layer graphene sheet (C338) is appended on the bottom side, giving a total composition Ti300C538O200. Atomic heights span z = 7.31–19.41 Å (thickness 12.10 Å) with 39.60 Å of vacuum along c. The graphene forms a narrow plane at z = 7.79 ± 0.31 Å (7.31–8.45 Å), the bottom *O plane is at z = 12.40 Å and the top *O plane at z = 19.41 Å; the graphene–MXene van der Waals gap is 4.61 Å (center-to-plane; minimum 3.95 Å to *O), and the graphene–bottom Ti plane separation is 6.78 Å. The result is a well-isolated graphene|Ti$_3$C$_2$O$_2$ heterojunction with symmetric *O terminations and a dense graphene areal carbon density of 41.49 nm^{-2}.

B4 (Cu-inserted MoS$_2$; same MoS$_2$ host as A3)

Orthorhombic P1 cell with a = 27.58 Å, b = 22.29 Å, c = 28.13 Å (α = β = γ = 90.00°; in-plane area 6.15 nm^2). The slab contains Mo$_{64}$Cu5S136 and occupies z = 4.80–8.32 Å (thickness 3.51 Å) with 24.61 Å of vacuum. The host S–Mo–S tri-layer is preserved ($\langle$z$\rangle$: S(bottom) = 5.03 Å, Mo = 6.59 Å, S(top) = 8.15 Å; S–S spacing 3.12 Å; Mo–S offsets 1.55/1.56 Å), as in A3. At A3's Mo vacancies, five Cu centers lie within the chalcogen sandwich at z = 6.55–6.89 Å ($\langle$z$\rangle$ = 6.70 Å), i.e., $\sim$0.11 Å above the Mo plane and biased toward the top S layer; the areal Cu density is 0.81 nm^{-2}. Local coordination around Cu is consistent with insertion at chalcogen-vacancy–like sites: each Cu shows $\sim$4–5 short Cu–S contacts at 2.28–2.38 Å (mean of first shell $\approx$ 2.31 Å) with the next S neighbors near 2.95–3.06 Å, and $\sim$2–3 Cu–Mo contacts at 2.78–3.26 Å. This configuration provides a well-defined model to probe how subsurface/embedded Cu at vacancy motifs modulates CO$_2$ binding relative to the pristine A3 monolayer.

CO$_2$ Sensor Performance Evaluation

The development of sensitive and selective CO$_2$ sensors addresses critical needs in environmental monitoring, industrial process control, and indoor air quality management. To evaluate the CO$_2$ sensing capabilities of candidates proposed by the DR agents (Figure DL_5 1), we quantified binding energies using a systematic grid-based sampling approach that captures the heterogeneity of surface adsorption sites, with g-C$_3$N$_4$ serving as the established benchmark reference [52].

The comprehensive sampling methodology employed a 5×5 grid across each surface unit cell, positioning CO$_2$ molecules at 25 distinct initial configurations to explore the complete adsorption landscape (Figure DL_5 2). At each grid point, CO$_2$ molecules were placed 1.8 Å above the highest surface atom within a 3.0 Å search radius, with initial linear configurations parallel to the surface. This systematic approach ensures unbiased sampling of potential binding sites, including atop atoms, bridge sites, hollow sites, and defect locations, providing statistically robust binding energy distributions rather than single-point estimates.

The resulting binding energy distributions reveal significant variability both within and between candidate systems (Figure DL_5 3). The violin plots demonstrate that surface heterogeneity plays a crucial role in CO$_2$ adsorption, with some candidates exhibiting narrow distributions indicating uniform binding sites, while others show broad distributions suggesting diverse adsorption environments. Mean binding energies range from approximately 0.15 to 0.90 eV across the evaluated systems, with A-series candidates generally demonstrating stronger CO$_2$ interactions compared to B-series materials.

Performance scoring using the linear transformation framework with sensitivity parameter $\beta = 0.2$ (Figure DL_5 4) quantifies the relative sensing capabilities normalized to the g-C$_3$N$_4$ baseline. The distribution of scores reveals that commercial DR agent successfully identified materials spanning a range of binding strengths, from weak physisorption suitable for rapid-response sensors to stronger chemisorption appropriate for CO$_2$ capture applications.

The systematic evaluation highlights the importance of comprehensive surface sampling in sensor material assessment. Candidates with both high mean binding energies and narrow distributions, such as B4, represent ideal sensor materials offering consistent response across the entire surface. Conversely, materials with broad binding energy distributions may exhibit non-uniform sensor responses, potentially complicating calibration and reducing reliability. These findings underscore that effective CO$_2$ sensor design requires not only optimizing average binding strength but also minimizing surface heterogeneity to ensure reproducible sensing performance.

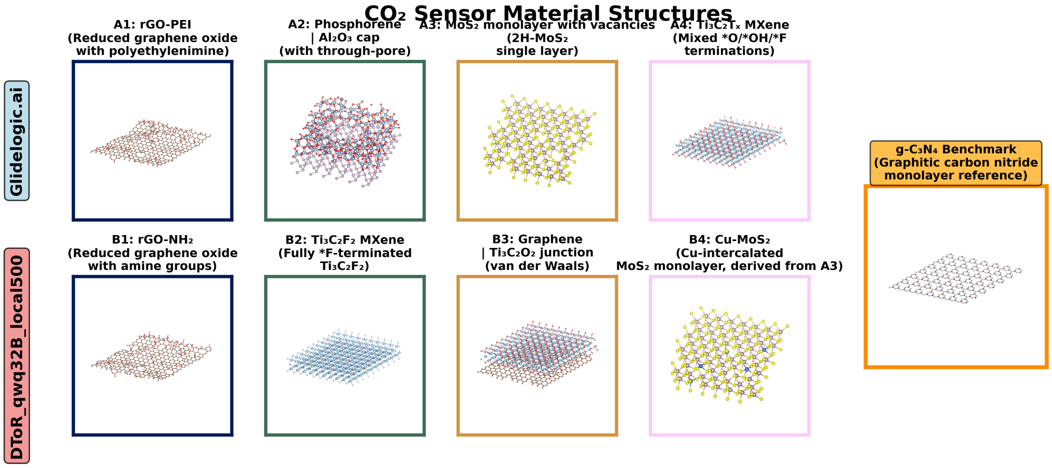


Figure DL_5 1: Structural configurations of CO_2 sensor candidates proposed by DR agents, with g-C_3N_4 [52] included as the domain benchmark reference.

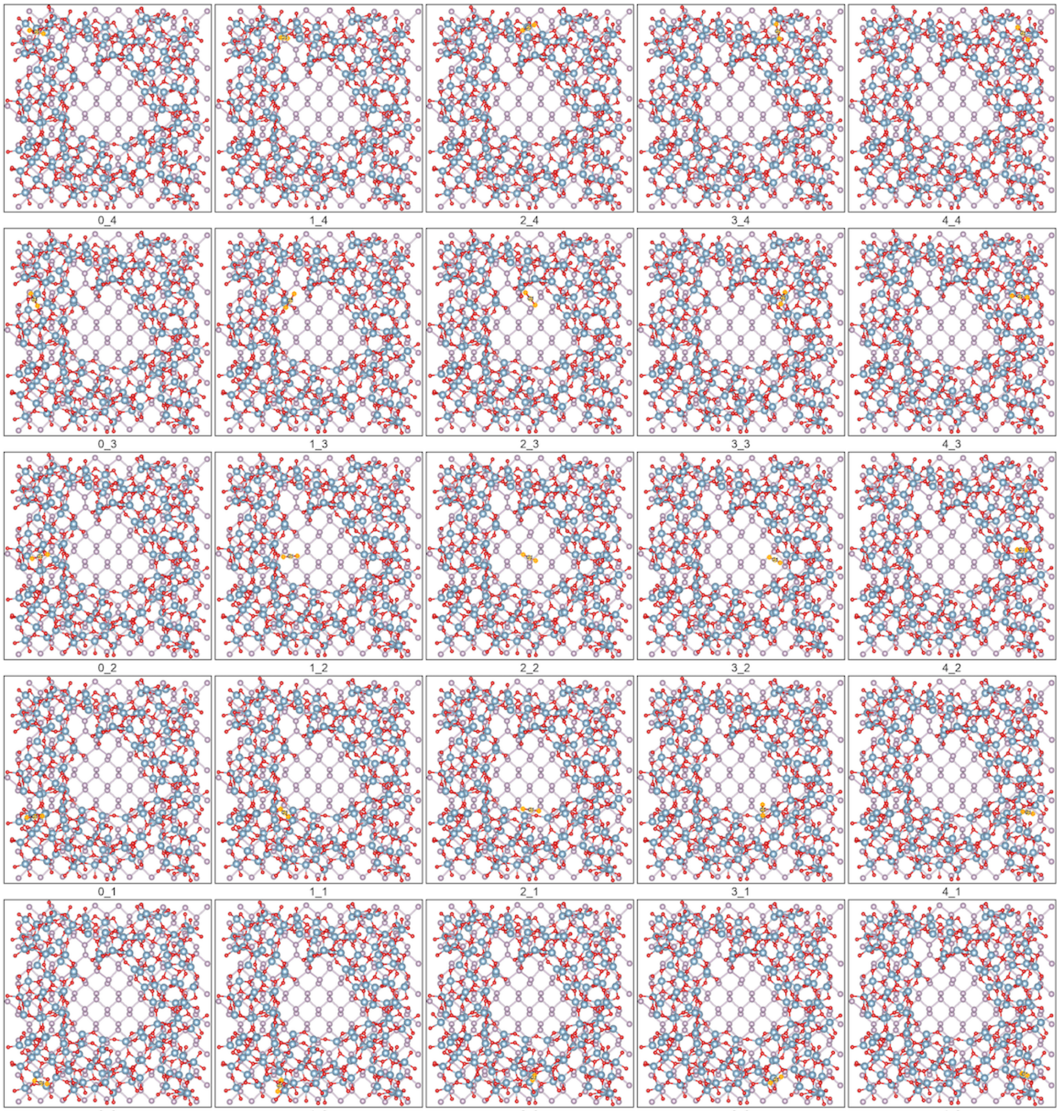


Figure DL_5 2: Top-view representation of the 5×5 grid sampling approach for CO_2 adsorption sites. Relaxed configurations show CO_2 molecules at 25 distinct initial positions on candidate A2, demonstrating comprehensive exploration of the binding landscape.

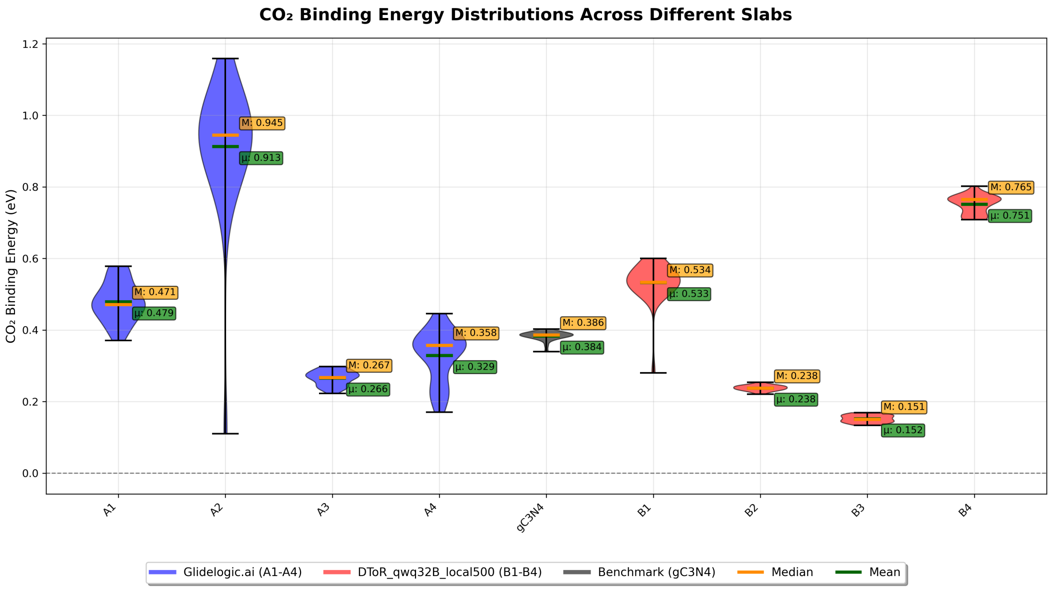


Figure DL_5 3: Distribution of CO_2 binding energies across all candidate systems visualized as violin plots. Mean (solid line) and median (dashed line) values are shown for each candidate, derived from 25 configurations per surface using the 5×5 grid sampling methodology.

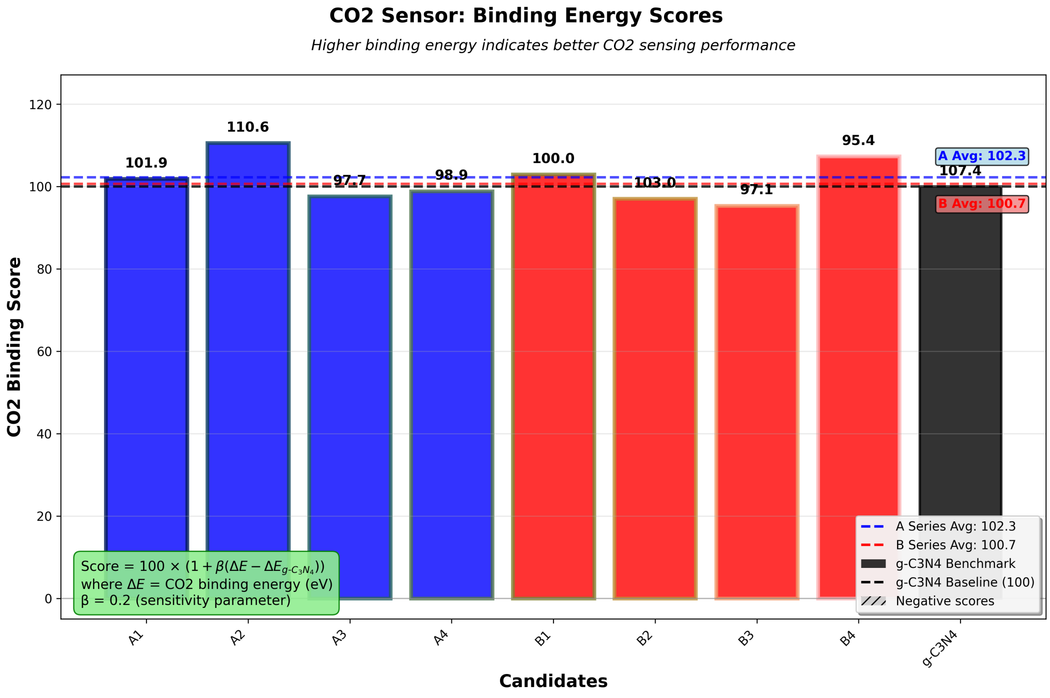


Figure DL_5 4: CO_2 sensing performance scores for all candidate systems from static DFT calculations, normalized to g-C_3N_4 baseline (100). Scores are calculated using the linear transformation with sensitivity parameter $\beta = 0.2$, where higher binding energies yield higher scores.

F Power Consumption Estimation

Power estimation and LLM/RAG time decomposition. We convert runtime to energy using a calibrated two–level power model on our workstation (**4×NVIDIA RTX A5500**, Ubuntu). System power was monitored continuously during experiments using `nvidia-smi` for GPU power draw and `powertop` for system–level consumption. We acknowledge that our power measurement methodology is relatively basic; however, the core contribution of this work lies in demonstrating the feasibility of edge–device deployment, which already showcases significant democratization and decentralization advantages over commercial subscription–based solutions. The quantitative power estimates provided here serve primarily as reference points for relative comparisons rather than definitive consumption benchmarks. Across runs we *observed* a typical baseline draw during retrieval/IO of $P_{\text{base}} \approx 76 \pm 3$ W (CPU, memory, disks, and idle GPUs). When the LLM generates tokens the GPU adds a model–dependent draw reported by `nvidia-smi`:

$$\Delta P_{\text{GPU}}(\text{ds14B}) = 282 \pm 4\ \text{W}, \quad \Delta P_{\text{GPU}}(\text{ds70B}) = 310 \pm 5\ \text{W}, \quad \Delta P_{\text{GPU}}(\text{qwq32B}) = 301 \pm 4\ \text{W}.$$

Hence $P_{\text{LLM}}(\text{model}) = P_{\text{base}} + \Delta P_{\text{GPU}}(\text{model})$ is 358 W (ds14B), 386 W (ds70B), and 377 W (qwq32B). During RAG (vector search over the local store, parsing, and file IO) the GPU utilization remained near–idle, so we bill RAG at $P_{\text{RAG}} = P_{\text{base}} \approx 76$ W.[4]

Empirical split of runtime into LLM vs. RAG. Let $\bar{t}_{\text{tot}}(s, m, k)$ be the mean wall–clock time for strategy $s \in \{\text{Single}, \text{DToR}$, model $m \in \{\text{ds14B}, \text{qwq32B}, \text{ds70B}\}$, and local–RAG size $k \in \{0, 100, 500\}$. For $k = 0$ (no local vector store) we roughly estimated that all time is attributable to LLM generation:

$$t_{\text{LLM}}(s, m, 0) = \bar{t}_{\text{tot}}(s, m, 0), \qquad t_{\text{RAG}}(s, m, 0) = 0.$$

For $k > 0$ we roughly split $\bar{t}_{\text{tot}}(s, m, k)$ into LLM and RAG by scaling the $k = 0$ LLM time to reflect the additional *LLM* passes used in local–knowledge summarization and complementary–query generation. The scale factors were estimated from repeated single–node micro–benchmarks (profiling isolated summarization and query–generation calls with coarse-grained timing) and calibrated against observed aggregate runtime patterns across full runs. Given measurement noise and workload variability, we report representative values that *differed by strategy*:

$$\begin{aligned}
\text{Single:}\quad & \alpha^{\text{Single}}_{\text{ds14B}} \approx 1.05, \quad \alpha^{\text{Single}}_{\text{qwq32B}} \approx 1.07, \quad \alpha^{\text{Single}}_{\text{ds70B}} \approx 1.09;\\
\text{DToR:}\quad & \alpha^{\text{DToR}}_{\text{ds14B}} \approx 1.10, \quad \alpha^{\text{DToR}}_{\text{qwq32B}} \approx 1.20, \quad \alpha^{\text{DToR}}_{\text{ds70B}} \approx 1.35.
\end{aligned}$$

We then compute

$$t_{\text{LLM}}(s, m, k) = \alpha^{s}_{m}\, \bar{t}_{\text{tot}}(s, m, 0), \qquad t_{\text{RAG}}(s, m, k) = \bar{t}_{\text{tot}}(s, m, k) - t_{\text{LLM}}(s, m, k) \ \text{(clipped at } \geq 0).$$

Two empirical patterns are worth noting. (i) In **Single** mode the increase in t_{LLM} from $k = 0$ to $k > 0$ is modest (ranging from $+5\%$ to $+9\%$ across models), and t_{RAG} exhibits a smaller difference than theoretically expected between $k = 100$ and $k = 500$ within measurement tolerance. (ii) In **DToR** mode, multi–hour runs accumulate more local–knowledge summarization and complementary–query steps; correspondingly t_{LLM} shows larger scaling (1.10/1.20/1.35 as above), and t_{RAG} increases with k due to heavier vector–store traffic and post–processing over longer context. When attribution is ambiguous we conservatively charge time to the LLM bucket.

Energy calculation. Per–report energy is the sum of LLM and RAG contributions:

$$E_{\text{LLM}}(s, m, k) = \frac{P_{\text{LLM}}(m)\, t_{\text{LLM}}(s, m, k)}{1000}, \quad E_{\text{RAG}}(s, m, k) = \frac{P_{\text{RAG}}\, t_{\text{RAG}}(s, m, k)}{1000}, \quad E_{\text{tot}} = E_{\text{LLM}} + E_{\text{RAG}} \ \ (\text{kWh}).$$

All results in Fig. 3d use $P_{\text{base}} = 76$ W, $\Delta P_{\text{GPU}} = \{282, 301, 310\}$ W (ds14B, qwq32B, ds70B), and the strategy–specific α values above. For example, substituting the measured $\{t_{\text{LLM}}, t_{\text{RAG}}\}$ of `DToR_qwq32B_local500` yields $E_{\text{tot}} = 4.16$ kWh per completed report; configuration–wise LLM/RAG splits are provided alongside runtime heatmaps in the Appendix.

---

[4] Short web fetches last seconds relative to multi–hour runs and are neglected.