

MIRA: Towards a General AI Scientist

Deep Principle Team

Artificial intelligence now reaches into much of scientific practice, yet a gap persists between systems that answer scientific questions and systems that advance science. Scientific discovery is not a single response but an iterative loop that runs from hypothesis generation through execution to interpretation, carried out under uncertainty and accountable to reproducible evidence. We present MIRA, an AI scientist tailored for the physical sciences, to operationalize this loop end to end. MIRA is organized as a squad of agents with different harnesses and tools that mirror how a human research team divides and coordinates its work. It frames research directions and mines the published literature and a team’s internal data to propose grounded, testable hypotheses; it executes pre- or user-defined simulation workflows on cloud compute infrastructure; and it drives physical experiments through our in-house automated laboratory, the AI Materials Factory, carrying a hypothesis from plan to bench. MIRA consolidates results into calibrated conclusions, states its uncertainty explicitly, and compiles each completed project into a persistent memory and a maintained scientific wiki, so that reasoning, provenance, and negative results remain available to later work. On two external benchmarks MIRA attains the highest score at the lowest cost per task among seven agents compared, and leads four of five task categories on a specialized drug-discovery suite, exceeding agents purpose-built for that domain. This report details MIRA’s architecture and stage-by-stage design, and demonstrates it across chemistry, materials, and energy as a generalist that accelerates research and discovery while remaining reliable and auditable.

Date: August 3, 2026

Correspondence: Deep Principle Team

Use: <https://mira.deepprinciple.com>



Deep Principle

1 Introduction

Artificial intelligence now touches nearly every part of scientific research. Large language models (LLMs) and agentic systems draft literature reviews, propose hypotheses, write and run code, analyze data, and help compose papers, building on a decade in which AI has learned to fold proteins and run self-driving materials labs [1, 2, 3, 4]. What has changed with LLMs is the breadth of that reach, where one general-purpose interface will respond, fluently and plausibly, to almost any scientific prompt. Yet this breadth conceals a gap [5]. Benchmarks can establish whether a model answers hard scientific questions. They reveal, however, little about whether it can move science forward—and each time frontier models master another expert exam [6, 7, 8], the two capabilities drift further apart. Real discovery is not a one-shot answer but a chain of interdependent choices made under uncertainty: a researcher frames a question, proposes explanations, converts them into runnable experiments or simulations, reads ambiguous and sometimes contradictory evidence, and lets what is learned reshape the next move [9].

That iterative discovery loop, not the isolated answer, is what a capable AI scientist must master. It decomposes into three interlocking capabilities, each with its own standard and its own failure modes [5]. Hypotheses must be not only original but checkable, and concrete enough to test rather than merely admire. Execution must be dependable and, wherever possible, automated within

the hard limits of instruments, compute, and cost. Interpretation must survive uncertainty, stay faithful to how the evidence was produced, and separate what the data establish from what they leave open. Fusing these stages into a working cycle demands more: crediting the stage responsible for an outcome, carrying experience from one investigation into the next, telling a genuine finding from the rediscovery of a known result, and keeping the dead ends and provenance that let a claim be reproduced and audited. Together these amount to a demanding specification, not merely a target for measuring AI scientists, but a bar any intelligent system must clear in practice.

A growing class of AI scientists has begun to assemble these pieces into longer, semi-autonomous workflows spanning literature search, code execution, hypothesis generation, and analysis [10, 11, 12, 13, 14, 15]. The strongest produce transparent, reproducible artifacts, place specialist agents under a coordinating planner, and sustain hours-long runs whose scientific yield grows with the compute invested [12, 14]. Their progress affirms a simple point—that the unit of scientific effort is a trajectory of decisions and evidence, not a lone answer. But that progress has clustered largely in the life sciences and in open-ended, literature-grounded analysis [13, 14]. Chemistry and materials, where a hypothesis must ultimately confront a physical molecule, reaction, composite materials, or device, have drawn far less attention, even though it is where progress is counted in new compounds and catalysts that unlock new technologies.

We developed MIRA exactly in this territory, and for the whole discovery loop rather than any single link in it. Three design choices give MIRA this reach. First, rather than a single agent with fixed harness, MIRA runs as a squad of specialized agents that mirror how a research group works. A coordinating agent sets direction and delegates to agents for literature, hypothesis, simulation, experiment, and analysis, which exchange results and check one another’s work. Second, MIRA is connected to elastic cloud compute infrastructure, turning a user-defined simulation workflow into anything from a single calculation to thousands of parallel jobs while handling scheduling, execution, and aggregation on its own. Third, and most distinctively, MIRA is wired into our in-house automated laboratory, the AI Materials Factory, where it schedules real synthesis and characterization, so that predictions that survive simulation are tested on physical hardware, and the measurements return to the same loop. Together, MIRA orchestrates all ingredients that lead to scientific discovery. These choices are borne out under external evaluation. On ResearchClawBench, an end-to-end research benchmark, MIRA scores 17.51 across the Chemistry, Energy, and Materials domains against 14.60 of Codex, among seven other agents, and does so at \$0.67 per task, the lowest cost in the comparison [16]. On the drug-discovery suite of SciAgentArena it reaches 81.1%, leading four of five task categories and exceeding general-purpose agent (such as Claude Code) as well as ToolUniverse, Biomni, and ChemToolAgent, each purpose-built for chemical or biomedical work, by at least 11.9 points [17]. As scientific value compounds, it preserves the reasoning, provenance, and intermediate steps of every project as reusable “discovery episodes” [5] that remain the data foundation still missing from the field.

Looking further ahead, MIRA points toward a different scale for discovery. Where a single scientist can weigh only a few hypotheses at a time, an AI scientist can pursue many at once, learn as much from each failure as each success, and carry insight from one project into the next, so a lab’s experience compounds instead of walking out the door when a project ends. As this reasoning drives ever larger fleets of simulations and self-driving laboratories [3, 18], the months that today separate a question from a credible answer can begin to compress toward days, and design spaces far too vast to survey by hand become navigable. The goal is not to write human scientists out of the process but to widen what a single lab can explore while keeping human judgment on the decisions that matter and building safeguards into a process that must stay trustworthy as it accelerates. MIRA is built for discovery, to help scientists reach past the edge of the known and bring back what they unravel.

2 Design

The specification set out above—hypotheses that are checkable, execution that is dependable, interpretation that survives uncertainty, and a memory that carries experience forward—has to be met by architecture. The three stages ask different things of the machinery that serves them, and no single mechanism provides all of them. MIRA therefore separates the loop into three layers joined by explicit interfaces: an *agent squad* that organizes reasoning and the division of labor; an execution layer, made up of *MIRA Compute* and the *AI Materials Factory*, that turns a plan into calculations and physical measurements; and a consolidation layer of *memory and a scientific wiki* that converts finished work into reusable knowledge (Fig. 1).

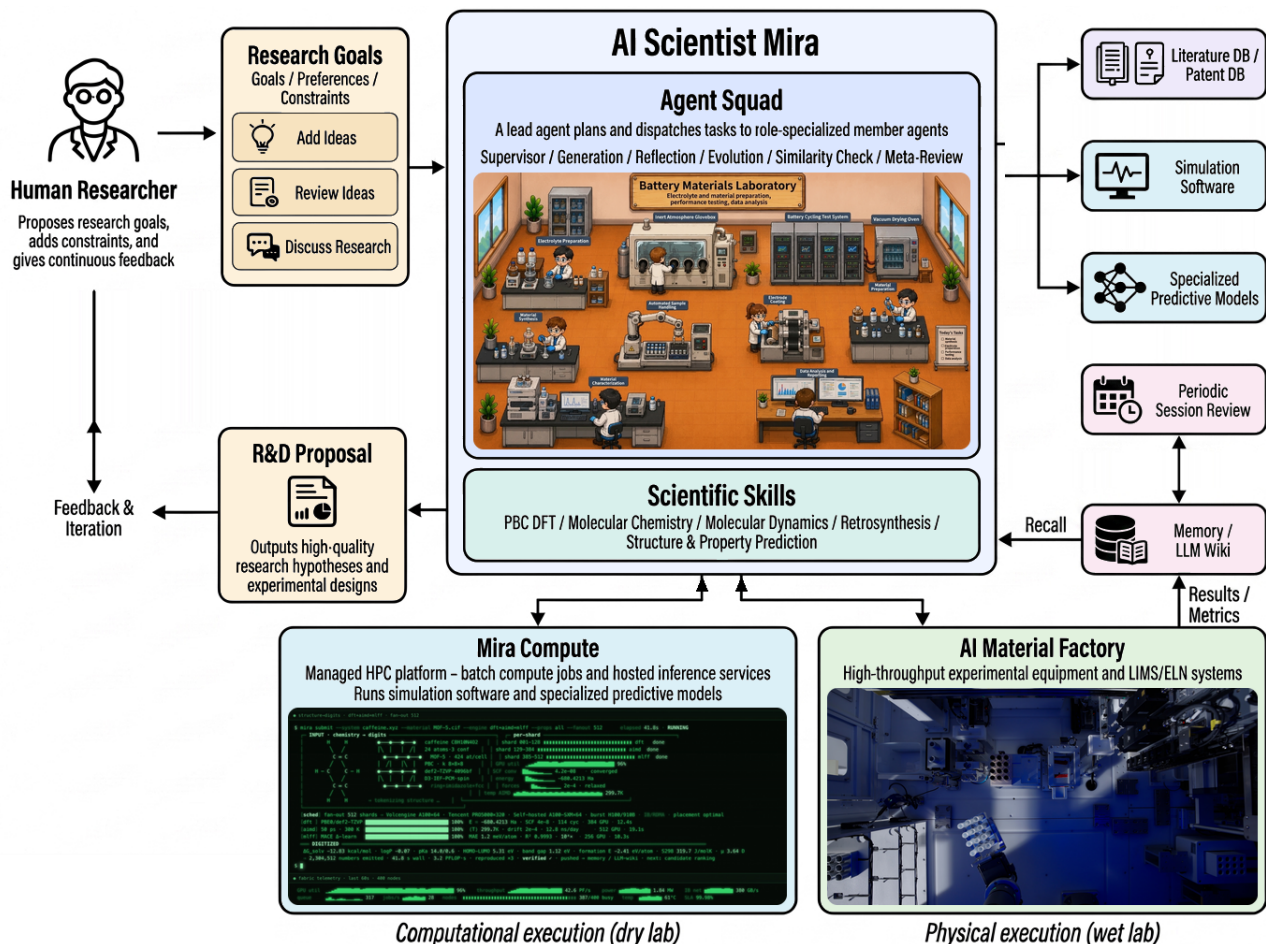


Fig. 1 | MIRA system architecture. An agent squad plans and delegates the work, drawing on scientific skills and executing through MIRA Compute and the AI Materials Factory; results are consolidated into memory and a scientific wiki and recalled in later runs.

2.1 Agent squad

The binding constraint on a single agent is its context window, which limits not only how much the agent can hold but how sharply it attends to what it is holding [19]. Consider a question that must be explored broadly—a dozen candidate mechanisms, each requiring several papers and a screening calculation. That is more material than one window holds, and well before the limit is reached the details that matter are competing for attention with everything else the agent carries. An agent forced

to compress drops alternatives instead of ranking them. A squad buys more than speed: several lines of inquiry stay open at once, each in full detail, and can then be compared, much as a research team actually works. Recent AI scientists have converged on the same conclusion, placing specialist agents under a coordinating planner [11, 14, 12].

MIRA’s squad follows the general agent-teams pattern, which rests on four elements [20]. A *lead* agent decomposes the goal, creates and assigns tasks, approves plans, and synthesizes what returns. *Member* agents each hold an independent context and carry a task to a deliverable. A *shared task board* lists the work items and their dependencies and is the single source of truth on what remains. A *mailbox* lets members address one another directly rather than routing everything through the lead. That last element is what separates a squad from a set of subagents: subagents report only to their caller and are blind to one another, while squad members share a board, can question one another’s findings, and can disagree.

The squad is staffed by role, not by topic, so that each member is accountable for a distinct scientific function. A supervising agent parses the goal and sets the rules of the run. Generative roles explore the literature and internal data to propose hypotheses, drawing on the internal knowledge of the underlying model as well as retrieved evidence [21]; reflective roles act as peer reviewers, checking correctness, novelty, and whether a proposal is simulable at all. An evolutionary role improves the survivors by simplification, recombination, and extension [15], while a similarity-checking role clusters near-duplicates so that a crowd of agents does not keep proposing variants of one idea. A synthesizing role produces the research overview and the global feedback that shapes the next round, and simulation and experiment roles carry what survives into the execution layer. This decomposition is close to the one validated in Co-Scientist [11], and the convergence suggests that the division of scientific labor among agents is settling into a stable design.

The squad is also a device for manufacturing disagreement. Sequential exploration suffers a well-known failure: once an agent finds one plausible explanation, subsequent search bends toward it and effectively stops. MIRA counters this by assigning members competing hypotheses and making refutation of the others part of their objective [22], so that an explanation survives only after the rest have tried to break it. Ranking escalates the same principle into a tournament, where hypotheses are compared pairwise and improved variants re-enter the competition [11], and evaluation does not stop at the first acceptable answer.

2.2 Scientific skills

Between the agents and the machines that execute their plans sits a layer of curated procedural knowledge, packaged as skills. A skill encodes what a competent practitioner of a workflow knows and a general-purpose model does not: which parameters a method requires, which convergence criteria make a result trustworthy, which pre-screening step should precede an expensive calculation, how raw output becomes a physical quantity, and what a valid result looks like. Skills range from computational workflows on periodic and molecular systems, molecular dynamics, retrosynthesis [23], structure prediction, and property prediction, to experimental workflows for chemical and materials synthesis, characterization, and purification. Other skills write the execution trajectory to our internal database and laboratory information systems, and synthesize the results.

Several of these skills are served by models we developed in house rather than by general-purpose tools, and the choice of model is part of what the skill encodes. Transition-state search is the clearest case, because it dominates the cost of reaction-network exploration and is the step most likely to fail silently. Rather than searching each saddle point from scratch, MIRA predicts it: OA-ReactDiff generates transition-state geometries directly from reactant and product structures with an object-aware equivariant diffusion model [24], and React-OT replaces stochastic sampling

with a deterministic optimal-transport map, which removes the variance of repeated sampling and lowers the cost of each structure [25]. Reactive machine-learned interatomic potentials, trained on a large database of molecular Hessians [26], then screen and refine the predicted structures so that first-principles calculations are committed only to candidates likely to converge [27], and reaction enumeration and network growth follow the YARP methodology [28]. Property prediction is served by a foundation model trained to generalize across experimental measurements rather than to fit a single property [29], and for inorganic design the generative roles draw on the internal chemical knowledge of large language models to propose functional metal complexes [21]. Because each of these is bound into a versioned skill, an improvement to any one of them propagates to every project that invokes it, without the agents having to know that it changed.

The layer exists because dependability at the execution stage cannot come from the reasoning layer. Asking a model to compose an input file from first principles on every run makes correctness a matter of that run’s luck; binding the workflow into a versioned skill makes it a property of the system. It also makes runs comparable: two projects that invoke the same skill have performed the same procedure, a precondition for the provenance and auditability the loop is meant to preserve.

2.3 MIRA Compute

MIRA Compute is the computational arm of the execution layer.

It is *cloud-native*: an elastic, multi-tenant scheduling plane over pooled heterogeneous resources. A workflow is submitted declaratively rather than as a batch script on a machine the user logs into, and the software environment of each calculation class is a versioned image rather than a hand-maintained installation. This is what lets a single user-defined workflow expand into thousands of parallel jobs, with scheduling, execution, and aggregation handled by the platform, and it places MIRA on the same trajectory as the wider move of scientific computing onto cloud-native infrastructure [30].

It is also *AI-native*: the consumer of its interface is an agent, not a person. Conventional computing portals encode a human operator into every assumption: someone who watches a queue, reads a log to decide what went wrong, judges whether to resubmit, and holds credentials. Replace that operator with an agent and each assumption becomes a defect. An agent-facing platform must therefore make submission idempotent, so that a repeated intention is not a repeated cost; make failure legible as structure rather than prose in a log, so that a system without human judgment can still act on it; and make results self-delivering, so that a finished job becomes evidence in the agent’s workspace rather than a file waiting to be fetched. Two further requirements are defensive: liveness must be explicit, so that running work is distinguishable from dead work without inference, and credentials must stay outside the agent’s reach, so that a component susceptible to prompt injection has nothing to leak [31]. Each requirement closes a gap that human judgment used to cover.

The platform exposes two kinds of work, which together span what physical-science research asks of computation. Batch jobs are asynchronous, queued, and artifact-producing: structure relaxation, electronic and vibrational properties, molecular dynamics, transition-state and migration-barrier searches [32], conformer sampling, thermodynamic properties, and machine-learned force-field pre-screening [27]. Hosted inference services are persistent endpoints serving specialized models for molecular property prediction [29], protein and complex structure prediction, and transition-state and reaction-network prediction [24, 25, 28]. A large model answering fast individual calls is the opposite profile and demands the opposite scheduling policy, so the two run on separate paths.

2.4 The AI Materials Factory

Simulation narrows a hypothesis space but does not close it, and in the physical sciences a claim is settled in the laboratory. The AI Materials Factory is MIRA’s physical execution arm: an in-house automated laboratory whose high-throughput synthesis and characterization equipment, together with its laboratory information systems, answers to the same squad that runs the calculations. Candidates—those that survive computational screening, and those generated under property guidance—are scheduled as real experiments. The measurements return to the loop that proposed them, so that the evidence settling a hypothesis and the reasoning that generated it stay part of one record. Autonomous laboratories have already shown that closed-loop synthesis and characterization are achievable [3, 18, 4, 33]. What the architecture adds is that the laboratory sits behind the same planning layer as the calculations—one of two execution backends, not a facility that receives handoffs. The agentic setup also makes new equipment easy to integrate, as long as it exposes an API.

2.5 Memory and scientific wiki

The dominant way of giving a model access to a corpus is retrieval: documents are indexed, fragments are fetched per query, and answers are generated [34]. Its structural weakness is that nothing accumulates. The model rediscovers what it needs on every question, and the cross-references, syntheses, and flagged conflicts produced by one analysis vanish when the session closes [35]. For science the loss is acute, because the most valuable knowledge a research program produces is the knowledge that never reaches the published literature: that a system does not converge under a given functional, that a specific parameterization is required to match experiment, that a synthesis route fails above a certain temperature. Such facts are purchased with machine time and reagents; they are what a departing researcher takes with them; and retrieval over papers offers nowhere to put them.

MIRA therefore compiles rather than retrieves. Raw sources—papers, patents, internal data, and the artifacts returned by execution—are immutable and never overwritten. Above them sits a maintained wiki of interlinked pages covering entities, methods, comparisons, and syntheses, carrying explicit cross-references and flagged contradictions. A schema governs how those pages are named, linked, and updated, so that maintenance is a defined procedure rather than an accumulation of habits. Three operations keep the layer alive. Ingestion compiles a new source or result into the wiki and propagates its consequences into the pages it affects. Querying draws its answers from the compiled layer with citations, then files the resulting analysis back as a page instead of letting it dissolve into conversation history. Periodic linting inspects the wiki for contradictions, for claims superseded by newer evidence, and for gaps, and it proposes what should be examined next [35].

Consolidation runs offline rather than in the moment. A periodic session review revisits completed sessions and their execution results, distills what merits retention into structured entries, writes them to memory and the wiki, and reconciles them with what is already recorded. Recall then injects the relevant fraction into later runs—the only path by which prior work reaches present reasoning. Running this while the system is otherwise idle, rather than during the interactions a researcher is waiting on, follows the same logic as sleep-time computation in agent systems [36]. It also has a scientific counterpart: the significance of a result is usually established after the bench work, when it is reconciled with what else is known.

This layer is what makes past work reusable rather than merely recorded [5]. A squad is by design volatile, since each member’s context ends with its run, and the artifacts of execution are inert, since output files do not become knowledge on their own. Compilation converts both into assets. Knowledge compounds across projects, so that a later investigation begins from established ground instead of re-deriving it. Negative results—the evidence most reliably lost in ordinary practice—acquire a home, so

that a failure recorded once need not be repeated. Conflicting results are recorded as conflicts rather than coexisting unnoticed, and that record is also what lets the system tell a genuine finding from the rediscovery of a known one, a distinction the loop cannot do without.

2.6 Closing the loop

The layers produce a discovery system only in combination. A research goal enters with its constraints and is decomposed by the squad, which proposes candidate hypotheses, prunes them by review and by similarity, and ranks what remains. Surviving candidates are dispatched to MIRA Compute, where cheap predictors screen broadly and expensive first-principles calculations adjudicate what passes; those that persist are scheduled in the AI Materials Factory, where they meet physical measurement. Results from both arms return to the squad, which verifies its own earlier reasoning against them and evolves the next round of hypotheses, and to the consolidation layer, which compiles the round, including what failed, into knowledge the next project recalls at its start. The squad supplies breadth and criticism, the execution layer converts hypotheses into evidence, and the consolidation layer turns that evidence into the priors of the next investigation.

3 Benchmark results

Benchmarks that score isolated answers provide limited evidence about a system’s capacity to conduct research [8]. We therefore evaluate MIRA on two benchmarks that assess complete research tasks rather than question answering: ResearchClawBench, which grounds each task in a published paper and scores the full trajectory from question to result, and the drug discovery domain of SciAgentArena, which scores an agent across the successive stages of a discovery workflow. On the former, MIRA obtains the highest score among seven agents at the lowest cost per task. On the latter, it obtains the highest score in four of five task categories and the highest minimum category score in the comparison.

3.1 Comparable research performance at lower cost

As MIRA is tailored for physical science, we select the Chemistry, Energy, and Materials domains from ResearchClawBench, an end-to-end research benchmark comprising 40 paper-grounded tasks across 10 scientific domains [16]. Each selected domain contains four tasks. We define a domain score as the arithmetic mean of its four task scores, and the overall score as the mean of the three domain scores. We run all seven agents in Figure 2, including MIRA, on GPT-5.4, so that the comparison reflects differences in agent harness and workloads rather than in the underlying model.

The three domains differ substantially in difficulty, and we observe the same ordering for every agent. Chemistry is the most difficult: no agent exceeds 10, and all seven scores fall within a 3.9-point interval between 6.00 and 9.90. Energy and Materials discriminate more strongly, spanning 9.8 and 7.4 points respectively. MIRA obtains the highest score in both of these domains, 23.30 in Energy and 19.79 in Materials, and the second-highest score in Chemistry, 9.45 against 9.90 for EvoScientist. Its overall score is 17.51 (Fig. 2a), followed by ResearchClaw at 15.60, EvoScientist at 15.00, Codex CLI at 14.57, OpenClaw at 13.63, ARIS Codex at 11.13, and Nanobot at 10.90. The margin between MIRA and the second-ranked agent is 1.91 points, which exceeds the 1.03-point interval separating the second- and fourth-ranked agents, and MIRA is the only agent to exceed 16.

For each agent we average the observed task cost within each selected domain and then average the three domain-level values, which are provider prices at the time of evaluation (Fig. 2b). MIRA operates at \$0.67 per task, the lowest cost in the comparison. Nanobot and OpenClaw are comparable in cost, at \$0.70 and \$0.72, but obtain substantially lower scores of 10.90 and 13.63. ResearchClaw, the

second-ranked agent by score, costs \$0.89. Higher expenditure does not recover the difference: Codex CLI at \$2.70 and EvoScientist at \$5.80 both score below MIRA, with EvoScientist costing 8.7 times as much per task for 2.5 points less. No agent in the comparison achieves either a lower cost or a higher score than MIRA, which therefore dominates all six alternatives on both criteria simultaneously.

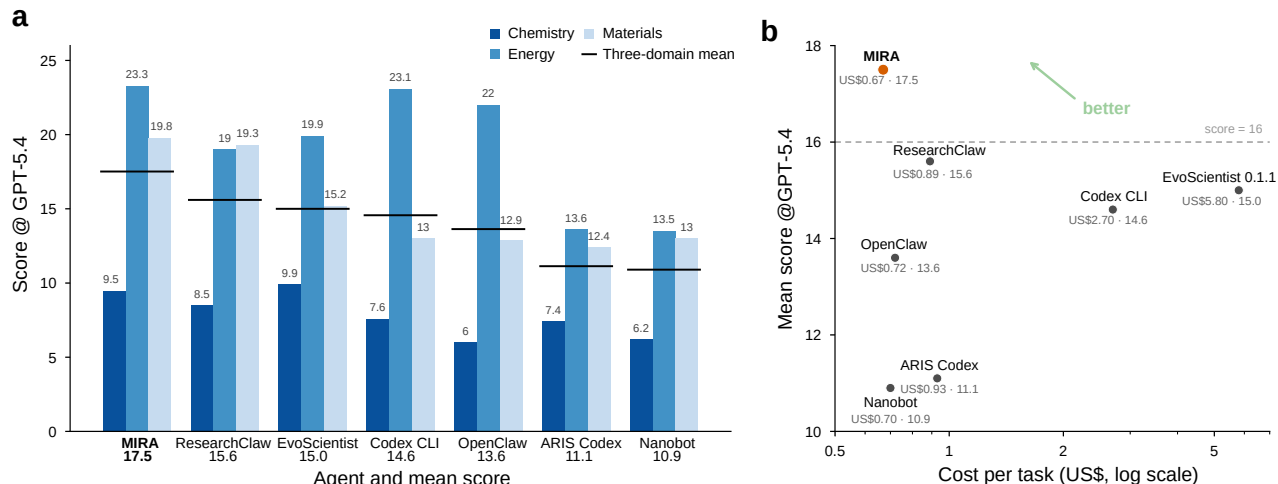


Fig. 2 | Cross-domain research quality and cost. a. Mean ResearchClawBench scores over the four tasks in each selected domain; horizontal markers show the unweighted three-domain mean, reported beneath each agent. **b.** Average task score against average task cost for the same agents. The logarithmic cost axis begins at \$0.50; labels report untransformed dollar values. Higher score and lower cost are preferable, so the upper-left region is favorable. MIRA attains both the highest score and the lowest cost, and is the only agent to exceed the 16-point score band.

3.2 Specialized drug-discovery performance

A cross-domain average primarily measures breadth. To assess whether a general-purpose system remains competitive within a single specialized domain, we evaluate MIRA on the drug discovery domain of SciAgentArena, a multi-institutional benchmark[17]. The domain comprises five task categories: chemical data preprocessing, chemical data analysis, molecule optimization, safety assessment, and claim validation. The SciAgentArena project lead supplied the comparison-agent results reported in Table 1; we evaluate MIRA on the same five categories and report the arithmetic mean of the five category scores as its overall score.

MIRA obtains an overall score of 81.1%, compared with 76.4% for Claude Code, 69.2% for ToolUniverse [37], 60.1% for Biomni [13], and 56.3% for ChemToolAgent [38]. It obtains the highest score in four of the five categories: data analysis (83.0), molecule optimization (78.0), safety assessment (84.0), and claim validation (78.5). The exception is data preprocessing, where ToolUniverse obtains 90.6 and Claude Code 86.6, against 82.0 for MIRA. Three of the four comparison agents, ToolUniverse, Biomni, and ChemToolAgent, are purpose-built for chemical and biomedical applications; MIRA exceeds all three by at least 11.9 points despite being a general-purpose system.

The distribution of scores across categories is as informative as the mean. MIRA’s five category scores lie between 78.0 and 84.0, a range of 6.0 points. The corresponding minima for the other agents are 71.8 for Claude Code, 52.0 for ToolUniverse, 48.5 for Biomni, and 44.9 for ChemToolAgent. ToolUniverse illustrates the pattern most clearly: it obtains the highest score in data preprocessing but falls to 52.0 in molecule optimization and 54.7 in safety assessment, indicating strong performance on data-handling tasks and weak performance on tasks that require chemical judgment. Because a

drug-discovery campaign requires all five stages in sequence, the minimum category score constrains end-to-end utility, making MIRA the most competitive drug discovery agents even though it is not designed for this subdomain.

Table 1. Performance in the SciAgentArena drug discovery domain. Scores are percentages. Overall is the arithmetic mean of the five task-category scores; bold values indicate the highest score in each column.

Agent	Data preprocessing	Data analysis	Molecule optimization	Safety assessment	Claim validation	Overall
MIRA	82.0	83.0	78.0	84.0	78.5	81.1
Claude Code	86.6	71.8	72.5	75.8	75.3	76.4
ToolUniverse	90.6	81.6	52.0	54.7	67.0	69.2
Biomni	77.6	48.5	59.5	55.6	59.2	60.1
ChemToolAgent	57.7	44.9	64.5	64.3	50.1	56.3

4 Use cases

The benchmarks above measure MIRA on tasks that carry established reference results. The three cases that follow are complete research projects, each beginning from a question posed in natural language and ending in a validated result. In each, the agent squad assembles the workflow itself rather than executing a pipeline we specify in advance. Drawing on its internal knowledge and harness, on memory, and on the scientific wiki, the squad establishes what is already known and what must be computed, then identifies the relevant scientific skills, selects the methods and tools those skills encode, sets the parameters, orders the stages, and dispatches the work to MIRA Compute. The cases span three distinct computational regimes: automated reaction-network exploration for a molecular system, multi-candidate structure prediction and interaction analysis for a protein-ligand complex, and periodic density-functional calculations for a metal surface. Each exercises a different set of skills, a different class of calculation, and a different form of validation, and each retains its intermediate structures, energies, and provenance in the project workspace. All three run entirely on the computational arm of the execution layer. The AI Materials Factory is still evolving rapidly, and we will describe closed-loop campaigns that carry hypotheses through wet-lab synthesis and characterization in a later report.

4.1 Predicting competing decomposition pathways of KHP

Automated exploration of reaction networks can complement experiment by predicting which pathways are kinetically accessible, which products should be monitored, and how short-lived intermediates behave. γ -ketohydroperoxides (KHPs) are important intermediates in autoxidation and pre-ignition chemistry [40], but their low concentrations, limited stability, and competing decomposition channels make their reactivity difficult to characterize experimentally [39].

We asked MIRA to explore two layers of the decomposition network of 3-hydroperoxypropanal, a representative γ -KHP. MIRA confirmed the task specification, retrieved and verified the molecular structure through PubChem [41], and configured the exploration (Fig. 3a). Within each layer it enumerated candidate reactions [28], dispatched parallel transition-state searches [24, 25, 27] and intrinsic reaction coordinate (IRC) validation [42] through MIRA Compute, and ranked the validated reactions by network-growth activation barrier. It enumerated 78 candidate reactions in the first

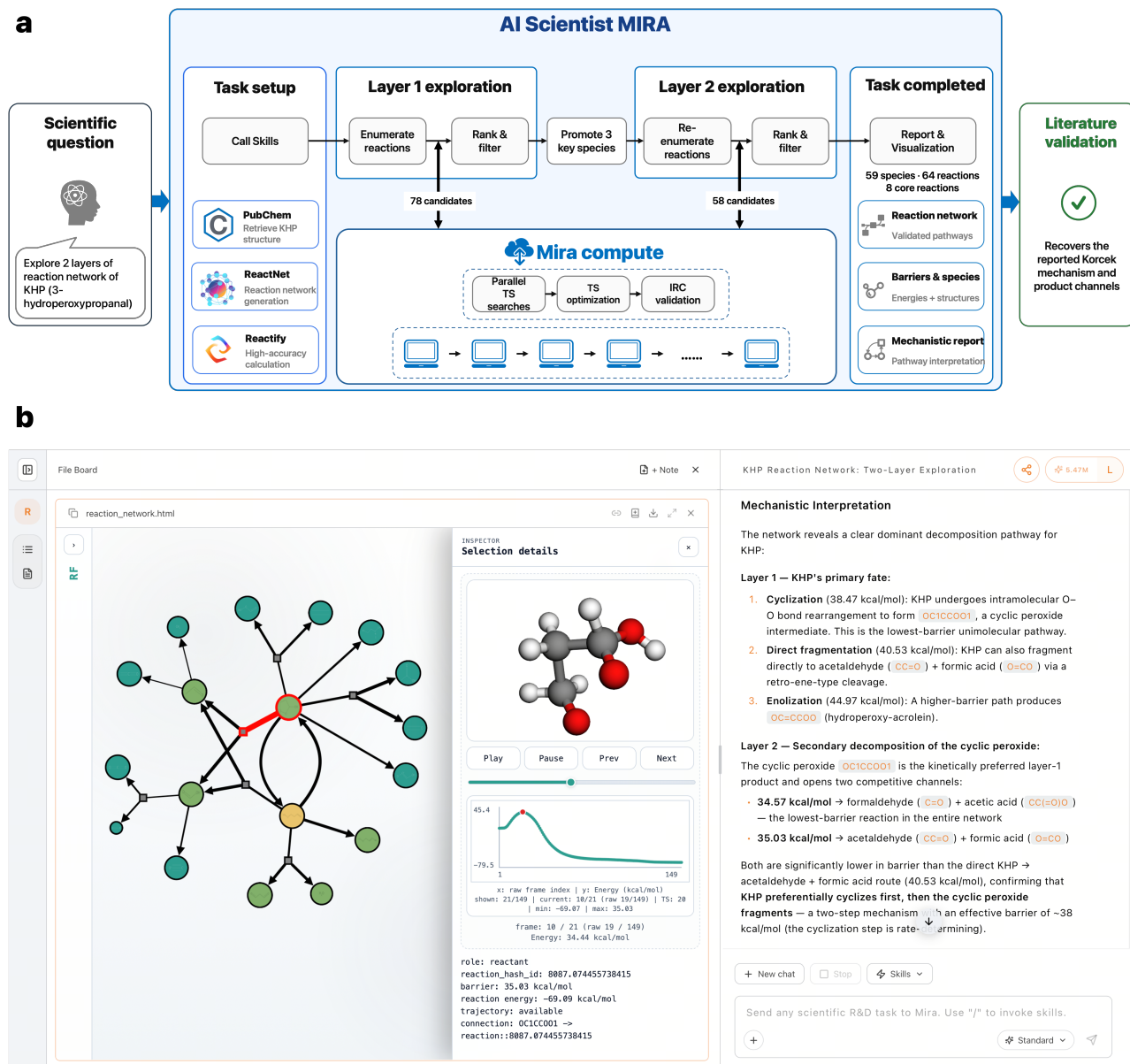


Fig. 3 | Autonomous exploration and interpretation of the KHP decomposition network. **a.** MIRA retrieves and verifies the KHP structure, configures a two-layer exploration, and dispatches candidate reactions to MIRA Compute for parallel transition-state searches, optimization, and IRC validation. The three most favorable species from the first layer are promoted into the second, yielding a final network of 59 species and 64 validated reactions, of which 8 are core reactions. **b.** The MIRA interface presents the validated reaction network, intermediate and transition-state structures, reaction-coordinate profiles, an interactive IRC trajectory, and the generated mechanistic report. The predicted pathways recover the cyclic-peroxide intermediate and the carbonyl and carboxylic acid product channels associated with the Korcek mechanism [39]. Session: <https://mira.deepprinciples.com/share/sessions/kSIIF3NQ>

layer, promoted the three most favorable species, and enumerated a further 58 candidates in the second. The completed network comprises 59 species connected by 64 validated reactions, of which 8 are core reactions. MIRA assembled the network, the corresponding intermediate and transition-state

structures, the reaction-coordinate profiles, and an interactive IRC viewer into a single workspace, together with a generated mechanistic report (Fig. 3b). Despite the complexity of the resulting network, these are delivered as interactive HTML files in which individual reactions and chemical species can be inspected directly.

The network resolves the competition between direct and stepwise decomposition. Direct fragmentation of KHP to acetaldehyde and formic acid proceeds over a barrier of 40.53 kcal/mol. Intramolecular cyclization to a five-membered cyclic peroxide is lower at 38.47 kcal/mol and is the lowest-barrier unimolecular pathway available to KHP. The cyclic peroxide then fragments over lower barriers still: 34.57 kcal/mol to formaldehyde and acetic acid, the lowest-barrier reaction in the network, and 35.03 kcal/mol to acetaldehyde and formic acid. Because the second step is more facile than the first, cyclization is rate-determining and the effective barrier of the two-step route is 38.47 kcal/mol, below the 40.53 kcal/mol required for direct fragmentation. MIRA therefore predicts that KHP cyclizes before it fragments. Both forward channels also lie well below the 45.87 kcal/mol barrier for reversion to KHP, so the cyclic peroxide is predicted to decompose faster than it reverts and is unlikely to accumulate. Enolization to hydroperoxy-acrolein at 44.97 kcal/mol and dehydration at 49.18 kcal/mol were also recovered, but neither is kinetically competitive with the cyclic route.

These predictions reproduce the defining features of the Korcek mechanism reported for this system, namely the five-membered cyclic peroxide intermediate and the carbonyl and carboxylic acid product channels [43, 39]. MIRA planned, executed, and interpreted the campaign from a single natural-language request, and returned both a mechanistic account and a set of experimentally testable predictions about which products to monitor.

4.2 Predicting and selecting ABL1–Imatinib binding modes

Protein–ligand complex structures provide the spatial context required to interpret molecular recognition and to identify the interactions that stabilize a small molecule within a binding site. When no experimental structure is available, computational prediction often returns several candidate binding modes with similar global confidence. Selecting among them is difficult, because a high global confidence score does not establish that the ligand orientation or the local interface is chemically reasonable. Candidates may share a protein fold while placing the ligand differently within the pocket, which produces different hydrogen-bond, π -stacking, ionic, and hydrophobic interaction patterns. A reliable workflow therefore has to evaluate both structural confidence and local binding interactions rather than select on a single score.

We asked MIRA to predict and analyze the complex formed by imatinib and the ABL1 kinase domain, a system whose binding mode has been resolved experimentally [44] and which therefore admits an independent check on the predicted interactions. From the ABL1 amino-acid sequence of 271 residues and the imatinib SMILES string [45], MIRA configured a two-stage workflow combining multi-candidate complex prediction, confidence-based ranking, structure preprocessing, and binding-mode analysis (Fig. 4a). It ran a multi-sample ESMFold2-fast prediction [46] on MIRA Compute, which returned three candidate complexes within a single run, together with structure files, confidence metrics, and model-quality outputs. MIRA ranked the candidates on mean predicted local distance difference test score (pLDDT), predicted template modeling score (pTM), and interface predicted template modeling score (ipTM) [2, 47], and carried all three into binding-mode analysis. It used Open Babel [48] for the format conversion required downstream and submitted the structures to ProteinsPlus, PoseEdit, and Protoss [49, 50, 51], which returned processed 3D complexes, 2D interaction records, and quantitative interaction summaries covering hydrogen bonds, hydrophobic contacts, and π -stacking interactions. MIRA assembled these outputs into a comparison board in which selecting a row of the ranking table switches the corresponding 2D diagram and 3D structure, and retained the

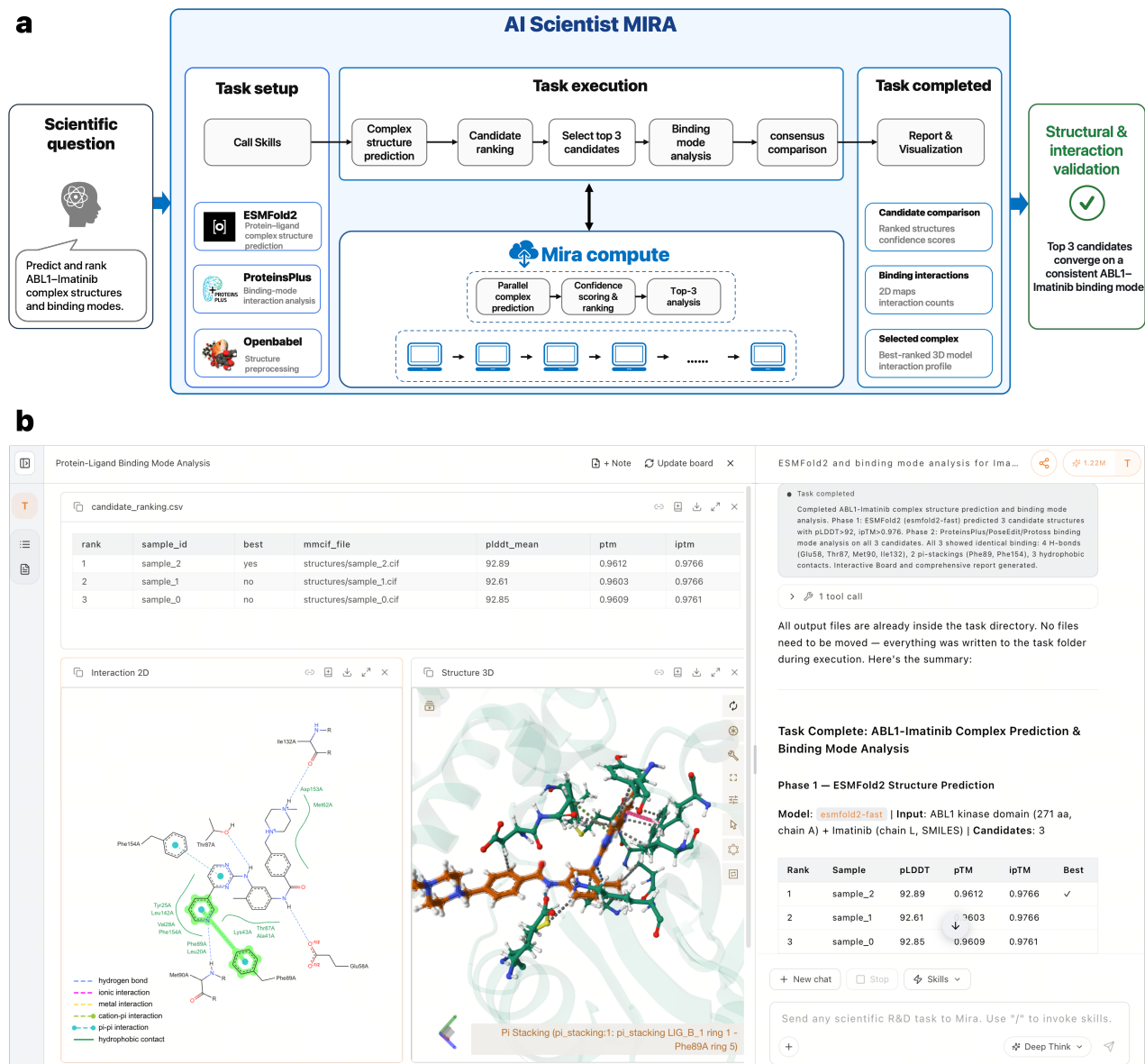


Fig. 4 | Prediction and selection of ABL1–Imatinib binding modes. **a.** MIRA calls ESMFold2 on MIRA Compute to generate and rank candidate complexes, then submits the ranked structures for binding-mode analysis with ProteinsPlus, PoseEdit, and Protoss. **b.** The MIRA interface presents the candidate ranking, 2D interaction maps, interactive 3D structures, and the generated report. The three candidates converge on a common binding mode, and **sample_2** is retained as the representative complex. Session: <https://mira.deepprinciple.com/share/sessions/6p0nHVTp>

underlying CSV, JSON, PDB, mmCIF, PAE, and distogram files in the project workspace (Fig. 4b).

The three candidates are closely spaced in confidence. Mean pLDDT ranges from 92.61 to 92.89, pTM from 0.9603 to 0.9612, and ipTM from 0.9761 to 0.9766, with **sample_2** ranking first at 92.89, 0.9612, and 0.9766. Because these margins are small relative to what such scores can resolve, MIRA did not treat the ranking alone as sufficient evidence for the predicted binding mode. It instead compared the protein–ligand interactions across all three candidates to determine whether structures sampled within one prediction run converge on a common interaction pattern. All three show the same four hydrogen bonds, to Glu58, Thr87, Met90, and Ile132, the same two π -stacking interactions, involving

Phe89 and Phe154, and three hydrophobic contacts. MIRA retained `sample_2` as the representative ABL1–Imatinib complex, with the agreement among the three separately sampled candidates serving as an internal consistency check on the shared binding mode. Because all three originate from the same model and the same prediction run, their agreement does not constitute independent cross-model or experimental validation.

Across this campaign MIRA configured multi-candidate prediction on MIRA Compute, ranked and selected structures, transferred them between specialized tools, coordinated the interaction analysis, and linked the ranking table to the 2D and 3D views. It retained the intermediate structures, confidence metrics, interaction records, and auxiliary model outputs in the project workspace, so that each step of the workflow remains available for inspection and reuse.

4.3 Modeling CO adsorption and charge redistribution on Pt(111)

Carbon monoxide adsorption on Pt(111) is a canonical system in surface science and heterogeneous catalysis, and a sensitive test of how an electronic-structure method describes adsorbate–surface bonding, site preference, and charge redistribution. The system is demanding because the calculated adsorption configuration and energy depend strongly on the treatment of exchange–correlation [53] and dispersion interactions. Experiment favors adsorption of CO atop a Pt atom, whereas several conventional density-functional approximations incorrectly stabilize hollow sites, a discrepancy known as the CO/Pt(111) puzzle [54]. Earlier calculations showed that including van der Waals interactions [55] recovers the observed atop preference and improves the predicted energetics [52]. Examining the adsorption energy together with the differential charge density gives complementary energetic and electronic views of the Pt–CO interaction.

We asked MIRA to calculate the adsorption energy of CO on Pt(111) and to analyze the associated charge redistribution (Fig. 5a). MIRA invoked pymatgen for bulk and surface-structure preparation, MACE for machine-learning-assisted geometry optimization, and ABACUS for DFT relaxation and self-consistent field calculations [56, 57, 58]. It prepared the bulk Pt structure, constructed and relaxed the Pt(111) slab, and relaxed the isolated CO molecule. Candidate adsorption configurations were then optimized by combining MACE–BFGS optimization with ABACUS DFT relaxation. For the selected configuration, MIRA ran three SCF calculations, on the adsorbed slab of 66 atoms, the bare slab of 64 atoms, and the isolated CO molecule, holding the simulation cell and real-space grid fixed across all three ($a = b = 11.213$ Å, $c = 26.867$ Å, $\gamma = 60^\circ$, a 100 Ry wavefunction cutoff, and a $135 \times 135 \times 324$ FFT grid). We define the adsorption energy as

$$E_{\text{ads}} = E_{\text{slab+CO}} - E_{\text{slab}} - E_{\text{CO}}, \quad (1)$$

and the differential charge density as

$$\Delta\rho = \rho_{\text{slab+CO}} - \rho_{\text{slab}} - \rho_{\text{CO}}. \quad (2)$$

MIRA collected the relaxed structures, total energies, charge-density files, and interactive visualizations in the project workspace (Fig. 5b).

The three SCF calculations give $-211,801.569$ eV for the adsorbed system, $-211,211.412$ eV for the bare Pt(111) slab, and -588.340 eV for isolated CO, and therefore an adsorption energy of -1.817 eV. The negative value indicates exothermic adsorption. The optimized structure places CO atop a Pt atom, consistent with experiment and with dispersion-inclusive DFT calculations on this system [52], and the differential charge density resolves the accumulation and depletion of charge around the Pt–CO interface and within the adsorbed molecule. Across this campaign MIRA translated a natural-language surface-science question into a multistage workflow, coordinated a machine-learning potential with

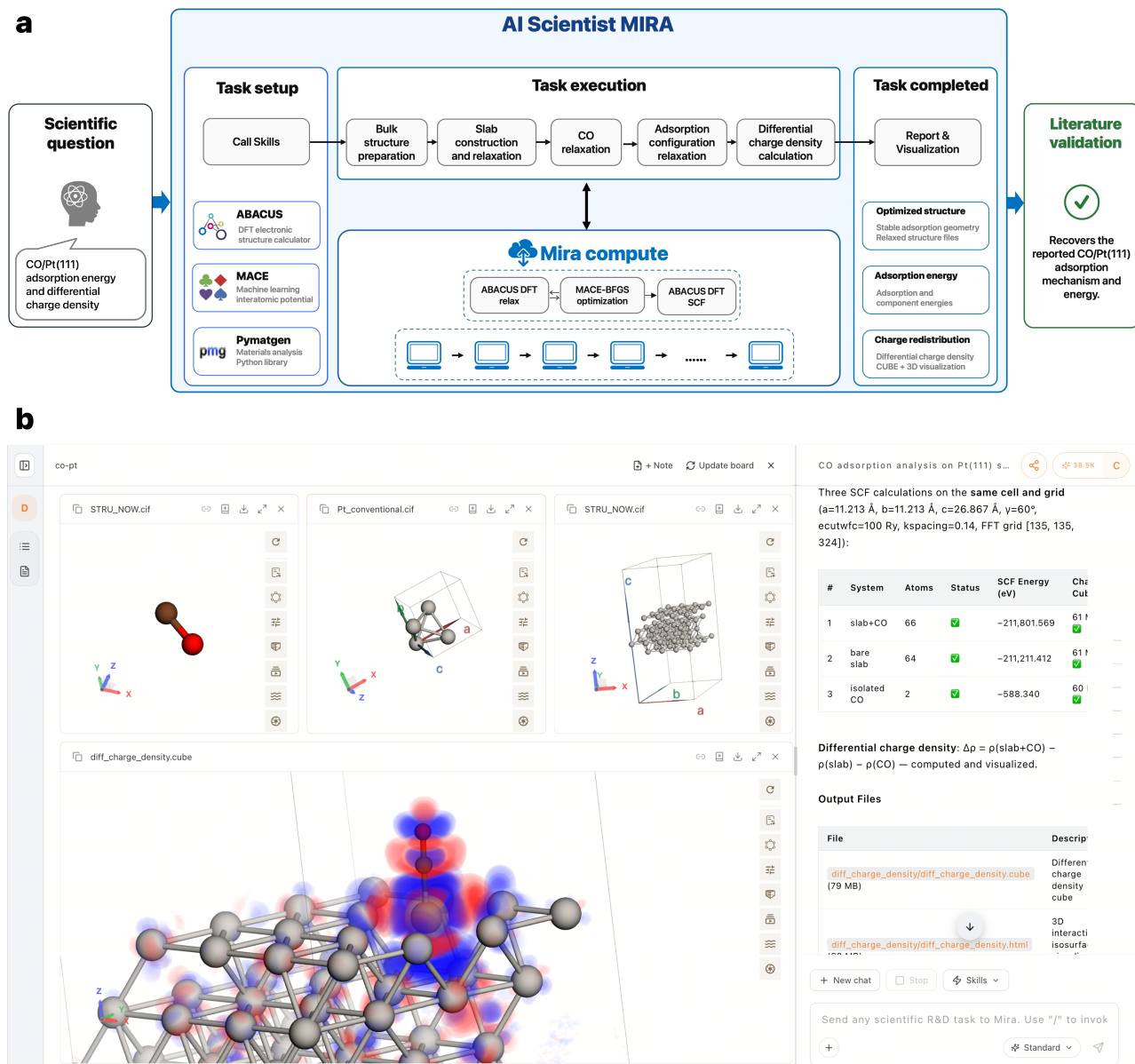


Fig. 5 | Modeling of CO adsorption and charge redistribution on Pt(111). **a.** MIRA prepares the bulk Pt structure and the Pt(111) slab, relaxes the isolated CO molecule and the adsorption configurations, and coordinates MACE-assisted optimization with ABACUS DFT relaxation and self-consistent-field calculations on MIRA Compute. Adsorption energies and differential charge densities are evaluated from consistently calculated adsorbed, bare-slab, and isolated-CO systems. **b.** The MIRA interface presents the optimized structures, the energy summary, and an interactive differential-charge-density visualization. The optimized configuration recovers the atop adsorption site reported for CO/Pt(111) with dispersion-inclusive functionals [52]. Session: <https://mira.deepprinciple.com/share/sessions/gn3vCNgZ>

first-principles calculations on MIRA Compute, held the cell and grid fixed across all three reference systems so that the energies remain directly comparable, and returned both the adsorption energetics and the electronic-structure visualization.

5 Outlook

The results reported here establish the core of our AI Scientist MIRA. On two external benchmarks MIRA attains the highest score at the lowest cost among the agents compared, and exceeds agents purpose-built for a single domain. The three case studies show the agent squad assembling and executing complete computational campaigns across molecular, biomolecular, and periodic materials systems, in each case from a natural-language question and without a workflow specified in advance. Taken together, these results establish the cognitive and execution layers of the architecture as a working foundation, and three developments now extend it.

The first is physical execution. The AI Materials Factory operates as MIRA’s experimental arm, addressable by the same squad that runs the calculations, and its agentic interface allows new instruments to be integrated as they come online. The facility is expanding rapidly, and closed-loop campaigns that carry a prediction from simulation through automated synthesis and characterization will be described in a forthcoming report. The second is iteration over longer horizons. The architecture returns computational and experimental results to the squad, which verifies its earlier reasoning against them and evolves the next generation of hypotheses. The campaigns presented above exercise this machinery within a single pass, and extending them into multi-round programs, in which each round is shaped by the outcome of the last, is the immediate direction of work and the setting in which the loop contributes most. The third is the accumulation of knowledge. Memory and the scientific wiki compile each campaign as it completes, so that reasoning, provenance, and negative results persist beyond the project that produced them. All three layers gain from use. Each completed campaign enlarges the compiled record that the squad recalls at the start of the next, refines the versioned skills through which the work is executed, and extends the range of instruments and workflows available to later projects, so that a research program of sufficient duration becomes progressively cheaper to extend. Quantifying that effect will require evaluation over a program rather than over an individual task, which remains an open problem [5] and one we intend to help address.

The longer-term direction is scale. Where an individual researcher can pursue only a small number of hypotheses concurrently, a system of this design can pursue many, and the binding constraint shifts from human attention to compute and instrument availability. As that constraint shifts, the months that now separate a question from a credible answer begin to compress toward days, and design spaces too large to survey by hand become navigable. Our aim is a system whose experience compounds rather than resets, where every result, including every failure, becomes ground for the next investigation, and where the reach of a researcher is set by the questions they choose to ask rather than by the hands available to test them.

6 Acknowledgement

We thank the entire Deep Principle team for discussion and support.

References

- [1] Wang, H., Fu, T., Du, Y., Gao, W., Huang, K., Liu, Z., Chandak, P., Liu, S., Van Katwyk, P., Deac, A., et al. Scientific discovery in the age of artificial intelligence. In: *Nature* 620.7972 (2023), pp. 47–60. URL: <https://doi.org/10.1038/s41586-023-06221-2>.

- [2] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with AlphaFold. In: *Nature* 596 (2021), pp. 583–589. URL: <https://doi.org/10.1038/s41586-021-03819-2>.
- [3] Szymanski, N. J., Rendy, B., Fei, Y., Kumar, R. E., He, T., Milsted, D., McDermott, M. J., Gallant, M., Cubuk, E. D., Merchant, A., Kim, H., Jain, A., Bartel, C. J., Persson, K., Zeng, Y. & Ceder, G. An autonomous laboratory for the accelerated synthesis of inorganic materials. In: *Nature* 624 (2023), pp. 86–91. URL: <https://doi.org/10.1038/s41586-023-06734-w>.
- [4] Burger, B., Maffettone, P. M., Gusev, V. V., Aitchison, C. M., Bai, Y., Wang, X., Li, X., Alston, B. M., Li, B., Clowes, R., Rankin, N., Harris, B., Sprick, R. S. & Cooper, A. I. A mobile robotic chemist. In: *Nature* 583.7815 (2020), pp. 237–241. URL: <https://doi.org/10.1038/s41586-020-2442-2>.
- [5] Du, Y., Dillmann, S., Laurent, J. M., Jansen, P., Jia, H., Schmidt, L., White, A. D., Persson, K. A., Arnold, F. H. & Duan, C. “Measuring AI Scientists: From Exams to Discovery”. Perspective. 2026.
- [6] Center for AI Safety, Scale AI & HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. In: *Nature* 649 (2026), pp. 1139–1146. URL: <https://doi.org/10.1038/s41586-025-09962-4>.
- [7] Wang, M., Lin, R., Hu, K., Jiao, J., Chowdhury, N., Chang, E. & Patwardhan, T. Frontier-Science: Evaluating AI’s Ability to Perform Expert-Level Scientific Tasks. In: *arXiv preprint arXiv:2601.21165* (2026). arXiv: 2601.21165 [cs.AI]. URL: <https://arxiv.org/abs/2601.21165>.
- [8] Song, Z., Lu, J., Du, Y., Yu, B., Pruyn, T. M., Huang, Y., Guo, K., Luo, X., Qu, Y., Qu, Y., Wang, Y., Wang, H., Guo, J., Gan, J., Shojaei, P., Luo, D., Bran, A. M., Li, G., Zhao, Q., Luo, S.-X. L., Zhang, Y., Zou, X., Zhao, W., Zhang, Y. F., Zhang, W., Zheng, S., Zhang, S., Khan, S. T., Rajabi-Kochi, M., Paradi-Maropakis, S., Baltoiu, T., Xie, F., Chen, T., Huang, K., Luo, W., Fang, M., Yang, X., Cheng, L., He, J., Hassoun, S., Zhang, X., Wang, W., Reddy, C. K., Zhang, C., Zheng, Z., Wang, M., Cong, L., Gomes, C. P., Hsieh, C.-Y., Nandy, A., Schwaller, P., Kulik, H. J., Jia, H., Sun, H., Moosavi, S. M. & Duan, C. Evaluating Large Language Models in Scientific Discovery. In: *arXiv preprint arXiv:2512.15567* (2025). arXiv: 2512.15567. URL: <https://arxiv.org/abs/2512.15567>.
- [9] Box, G. E. P. Science and statistics. In: *Journal of the American Statistical Association* 71.356 (1976), pp. 791–799. URL: <https://doi.org/10.1080/01621459.1976.10480949>.
- [10] Lu, C., Lu, C., Lange, R. T., Yamada, Y., Hu, S., Foerster, J., Ha, D. & Clune, J. Towards end-to-end automation of AI research. In: *Nature* 651 (2026), pp. 914–919. URL: <https://doi.org/10.1038/s41586-026-10265-5>.
- [11] Gottweis, J., Weng, W.-H., Daryin, A., Tu, T., et al. Accelerating scientific discovery with Co-Scientist. In: *Nature* (2026). URL: <https://doi.org/10.1038/s41586-026-10644-y>.
- [12] Anthropic. *Claude Science, an AI workbench for scientists*. 2026. URL: <https://www.anthropic.com/news/claude-science-ai-workbench>.
- [13] Huang, K., Zhang, S., Wang, H., Qu, Y., Lu, Y., Li, R., Roohani, Y., Qiu, L., Cao, S., Li, G., Zhang, J., Yin, D., Wierenga, R., Kavi, D., Liu, S., She, T., Marwaha, S., Carter, J. N., Zhou, X., Wheeler, M. T., Bernstein, J. A., Wang, M., He, P., Zhou, J., Snyder, M. P., Cong, L., Regev, A. & Leskovec, J. Autonomous biomedical research with an artificial intelligence agent. In: *Science* (2026), ead4351. URL: <https://www.science.org/doi/10.1126/science.adz4351>.

- [14] Mitchener, L., Yiu, A., Chang, B., Bourdenx, M., Nadolski, T., Sulovari, A., Landsness, E. C., Barabasi, D. L., Narayanan, S., Evans, N., Reddy, S., Foiani, M., Kamal, A., Shriver, L. P., Cao, F., Wassie, A. T., Laurent, J. M., Melville-Green, E., Caldas, M., Bou, A., Roberts, K. F., Zagorac, S., Orr, T. C., Orr, M. E., Zvezdaryk, K. J., Ghareeb, A. E., McCoy, L., Gomes, B., Ashley, E. A., Duff, K. E., Buonassisi, T., Rainforth, T., Bateman, R. J., Skarlinski, M., Rodriques, S. G., Hinks, M. M. & White, A. D. Kosmos: An AI Scientist for Autonomous Discovery. In: *arXiv preprint arXiv:2511.02824* (2025). arXiv: 2511.02824 [cs.AI]. URL: <https://arxiv.org/abs/2511.02824>.
- [15] Du, Y., Yu, B., Liu, T., Shen, T., Chen, J., Rittig, J. G., Sun, K., Zhang, Y., Song, Z., Zhou, B., Masschelein, C., Wang, Y., Wang, H., Jia, H., Zhang, C., Zhao, H., Ester, M., Head-Gordon, T., Gomes, C. P., Sun, H., Duan, C., Schwaller, P. & Jin, W. Accelerating Scientific Discovery with Autonomous Goal-evolving Agents. In: *arXiv preprint arXiv:2512.21782* (2025). arXiv: 2512.21782. URL: <https://arxiv.org/abs/2512.21782>.
- [16] Xu, W., Li, S., Ye, T., Cao, Q., Chen, Y., Gao, H., Wang, Y., Li, Q., Li, K., Xu, S., et al. ResearchClawBench: A Benchmark for End-to-End Autonomous Scientific Research. In: *arXiv preprint arXiv:2606.07591* (2026). URL: <https://arxiv.org/abs/2606.07591>.
- [17] Liu, T., Wang, A. X., Panescu, A., Chen, L. X., Long, W., Wei, X., Jing, Y., Zeng, Z., et al. Benchmarking AI Agents for Addressing Scientific Challenges Across Scales. In: *arXiv preprint arXiv:2606.12736* (2026). SciAgentArena. URL: <https://arxiv.org/abs/2606.12736>.
- [18] Boiko, D. A., MacKnight, R., Kline, B. & Gomes, G. Autonomous chemical research with large language models. In: *Nature* 624 (2023), pp. 570–578. URL: <https://doi.org/10.1038/s41586-023-06792-0>.
- [19] Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F. & Liang, P. Lost in the Middle: How Language Models Use Long Contexts. In: *Transactions of the Association for Computational Linguistics* 12 (2024), pp. 157–173. URL: https://doi.org/10.1162/tacl_a_00638.
- [20] Anthropic. *Orchestrate teams of Claude Code sessions*. 2026. URL: <https://code.claude.com/docs/en/agent-teams>.
- [21] Lü, J., Song, Z., Zhao, Q., Du, Y., Cao, Y., Jia, H. & Duan, C. Generative Design of Functional Metal Complexes Utilizing the Internal Knowledge and Reasoning Capability of Large Language Models. In: *Journal of the American Chemical Society* 147.36 (2025), pp. 32377–32388. URL: <https://doi.org/10.1021/jacs.5c02097>.
- [22] Du, Y., Li, S., Torralba, A., Tenenbaum, J. B. & Mordatch, I. Improving Factuality and Reasoning in Language Models through Multiagent Debate. In: *arXiv preprint arXiv:2305.14325* (2023). arXiv: 2305.14325. URL: <https://arxiv.org/abs/2305.14325>.
- [23] Segler, M. H. S., Preuss, M. & Waller, M. P. Planning chemical syntheses with deep neural networks and symbolic AI. In: *Nature* 555.7698 (2018), pp. 604–610. URL: <https://doi.org/10.1038/nature25978>.
- [24] Duan, C., Du, Y., Jia, H. & Kulik, H. J. Accurate transition state generation with an object-aware equivariant elementary reaction diffusion model. In: *Nature Computational Science* 3.12 (2023), pp. 1045–1055. URL: <https://doi.org/10.1038/s43588-023-00563-7>.
- [25] Duan, C., Liu, G.-H., Du, Y., Chen, T., Zhao, Q., Jia, H., Gomes, C. P., Theodorou, E. A. & Kulik, H. J. Optimal transport for generating transition states in chemical reactions. In: *Nature Machine Intelligence* 7 (2025), pp. 615–626. URL: <https://doi.org/10.1038/s42256-025-01010-0>.

- [26] Cui, T., Han, Y., Jia, H., Duan, C. & Zhao, Q. A Large Scale Molecular Hessian Database for Optimizing Reactive Machine Learning Interatomic Potentials. In: *Scientific Data* 13 (2025), p. 37. URL: <https://doi.org/10.1038/s41597-025-06350-5>.
- [27] Zhao, Q., Han, Y., Zhang, D., Wang, J., Zhong, P., Cui, T., Yin, B., Cao, Y., Jia, H. & Duan, C. Harnessing Machine Learning to Enhance Transition State Search with Interatomic Potentials and Generative Models. In: *Advanced Science* 12.34 (2025), e06240. URL: <https://doi.org/10.1002/advs.202506240>.
- [28] Zhao, Q. & Savoie, B. M. Simultaneously improving reaction coverage and computational cost in automated reaction prediction tasks. In: *Nature Computational Science* 1 (2021), pp. 479–490. URL: <https://doi.org/10.1038/s43588-021-00101-3>.
- [29] Deep Principle Team. *Materials Property Axiom: Scaling Foundation Models to Experimental Property Generalists via Multi-phase Training*. 2026. URL: <https://www.deepprinciple.com/papers/mpa.pdf>.
- [30] Kamatar, A., Pauloski, J. G., Babuji, Y., Chard, R., Sakarvadia, M., Babnigg, D., Chard, K. & Foster, I. Empowering Scientific Workflows with Federated Agents. In: *arXiv preprint arXiv:2505.05428* (2025). To appear at IEEE IPDPS '26. arXiv: 2505.05428 [cs.DC]. URL: <https://arxiv.org/abs/2505.05428>.
- [31] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T. & Fritz, M. Not what you’ve signed up for: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In: *arXiv preprint arXiv:2302.12173* (2023). arXiv: 2302.12173. URL: <https://arxiv.org/abs/2302.12173>.
- [32] Henkelman, G., Uberuaga, B. P. & Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. In: *The Journal of Chemical Physics* 113.22 (2000), pp. 9901–9904. URL: <https://doi.org/10.1063/1.1329672>.
- [33] MacLeod, B. P., Parlane, F. G. L., Morrissey, T. D., Häse, F., Roch, L. M., Dettelbach, K. E., Moreira, R., Yunker, L. P. E., Rooney, M. B., Deeth, J. R., Lai, V., Ng, G. J., Situ, H., Zhang, R. H., Elliott, M. S., Haley, T. H., Dvořák, D., Aspuru-Guzik, A., Hein, J. E. & Berlinguette, C. P. Self-driving laboratory for accelerated discovery of thin-film materials. In: *Science Advances* 6.20 (2020), eaaz8867. URL: <https://doi.org/10.1126/sciadv.aaz8867>.
- [34] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S. & Kiela, D. “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”. In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020. arXiv: 2005.11401. URL: <https://arxiv.org/abs/2005.11401>.
- [35] Karpathy, A. *LLM wiki*. Public gist. 2026. URL: <https://gist.github.com/karpathy/442a6bf555914893e9891c11519de94f>.
- [36] Lin, K., Snell, C., Wang, Y., Packer, C., Wooders, S., Stoica, I. & Gonzalez, J. E. Sleep-time Compute: Beyond Inference Scaling at Test-time. In: *arXiv preprint arXiv:2504.13171* (2025). arXiv: 2504.13171 [cs.AI]. URL: <https://arxiv.org/abs/2504.13171>.
- [37] Gao, S., Zhu, R., Sui, P., Kong, Z., Aldogom, S., Huang, Y., Noori, A., Shamji, R., Parvataneni, K., Tsiligkaridis, T. & Zitnik, M. Democratizing AI scientists using ToolUniverse. In: *arXiv preprint arXiv:2509.23426* (2025). arXiv: 2509.23426. URL: <https://arxiv.org/abs/2509.23426>.

- [38] Yu, B., Baker, F. N., Chen, Z., Herb, G., Gou, B., Adu-Ampratwum, D., Ning, X. & Sun, H. “ChemToolAgent: The Impact of Tools on Language Agents for Chemistry Problem Solving”. In: *Findings of the Association for Computational Linguistics: NAACL 2025*. 2025. arXiv: [2411.07228](https://arxiv.org/abs/2411.07228). URL: <https://arxiv.org/abs/2411.07228>.
- [39] Grambow, C. A., Jamal, A., Li, Y.-P., Green, W. H., Zádor, J. & Suleimanov, Y. V. Unimolecular Reaction Pathways of a γ -Ketohydroperoxide from Combined Application of Automated Reaction Discovery Methods. In: *Journal of the American Chemical Society* 140.3 (2018), pp. 1035–1048. URL: <https://doi.org/10.1021/jacs.7b11009>.
- [40] Zádor, J., Taatjes, C. A. & Fernandes, R. X. Kinetics of elementary reactions in low-temperature autoignition chemistry. In: *Progress in Energy and Combustion Science* 37.4 (2011), pp. 371–421. URL: <https://doi.org/10.1016/j.pecs.2010.06.006>.
- [41] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., Li, Q., Shoemaker, B. A., Thiessen, P. A., Yu, B., Zaslavsky, L., Zhang, J. & Bolton, E. E. PubChem 2025 update. In: *Nucleic Acids Research* 53.D1 (2025), pp. D1516–D1525. URL: <https://doi.org/10.1093/nar/gkaf1059>.
- [42] Fukui, K. The path of chemical reactions—the IRC approach. In: *Accounts of Chemical Research* 14.12 (1981), pp. 363–368. URL: <https://doi.org/10.1021/ar00072a001>.
- [43] Jalan, A., Alecu, I. M., Meana-Pañeda, R., Aguilera-Iparraguirre, J., Yang, K. R., Merchant, S. S., Truhlar, D. G. & Green, W. H. New Pathways for Formation of Acids and Carbonyl Products in Low-Temperature Oxidation: The Korcek Decomposition of γ -Ketohydroperoxides. In: *Journal of the American Chemical Society* 135.30 (2013), pp. 11100–11114. URL: <https://doi.org/10.1021/ja4034439>.
- [44] Schindler, T., Bornmann, W., Pellicena, P., Miller, W. T., Clarkson, B. & Kuriyan, J. Structural Mechanism for STI-571 Inhibition of Abelson Tyrosine Kinase. In: *Science* 289.5486 (2000), pp. 1938–1942. URL: <https://doi.org/10.1126/science.289.5486.1938>.
- [45] Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. In: *Journal of Chemical Information and Computer Sciences* 28.1 (1988), pp. 31–36. URL: <https://doi.org/10.1021/ci00057a005>.
- [46] Candido, S., Hayes, T., Derry, A., Rao, R., Lin, Z., Verkuil, R., Wu, B. Z., Lee, J. S., Bruguera, E. S., Keval, J. A., Kopylov, M., Pak, J. E., Wu, W., Thomas, N., Mataraso, S., Hsu, A., Trotman-Grant, A. C., Fatras, K., Costa, A. d. S., Badkundri, R., Akin, H., Oktay, D., Deaton, J., Montabana, E., Sitwala, H., Yu, Y., Wiggert, M., Carlin, D. A., Goering, A. W., Blazejewski, T., Sandora, M., Hla, M., Jia, T. Z., Kloker, L. H., Sofroniew, N. J., Uehara, M., Pannu, J., Bachas, S., Liu, D. S., Sercu, T. & Rives, A. Language Modeling Materializes a World Model of Protein Biology. In: *bioRxiv* (2026). URL: <https://doi.org/10.64898/2026.06.03.729735>.
- [47] Evans, R., O’Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Židek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zieliński, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., Kohli, P., Jumper, J. & Hassabis, D. Protein complex prediction with AlphaFold-Multimer. In: *bioRxiv* (2021). URL: <https://doi.org/10.1101/2021.10.04.463034>.
- [48] O’Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. & Hutchison, G. R. Open Babel: An open chemical toolbox. In: *Journal of Cheminformatics* 3.1 (2011), p. 33. URL: <https://doi.org/10.1186/1758-2946-3-33>.
- [49] Ehrt, C., Schulze, T., Graef, J., Diedrich, K., Pletzer-Zelgert, J. & Rarey, M. Proteins Plus: a publicly available resource for protein structure mining. In: *Nucleic Acids Research* 53.W1 (2025), W478–W484. URL: <https://doi.org/10.1093/nar/gkaf377>.

- [50] Diedrich, K., Krause, B., Berg, O. & Rarey, M. PoseEdit: enhanced ligand binding mode communication by interactive 2D diagrams. In: *Journal of Computer-Aided Molecular Design* 37.10 (2023), pp. 491–503. URL: <https://doi.org/10.1007/s10822-023-00522-4>.
- [51] Bietz, S., Urbaczek, S., Schulz, B. & Rarey, M. Protoss: a holistic approach to predict tautomers and protonation states in protein-ligand complexes. In: *Journal of Cheminformatics* 6.1 (2014), p. 12. URL: <https://doi.org/10.1186/1758-2946-6-12>.
- [52] G., L. K., Kundappaden, I. & Chatanathodi, R. A DFT study of CO adsorption on Pt(111) using van der Waals functionals. In: *Surface Science* 681 (2019), pp. 143–148. URL: <https://doi.org/10.1016/j.susc.2018.12.001>.
- [53] Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized Gradient Approximation Made Simple. In: *Physical Review Letters* 77.18 (1996), pp. 3865–3868. URL: <https://doi.org/10.1103/PhysRevLett.77.3865>.
- [54] Feibelman, P. J., Hammer, B., Nørskov, J. K., Wagner, F., Scheffler, M., Stumpf, R., Watwe, R. & Dumesic, J. The CO/Pt(111) Puzzle. In: *The Journal of Physical Chemistry B* 105.18 (2001), pp. 4018–4025. URL: <https://doi.org/10.1021/jp002302t>.
- [55] Dion, M., Rydberg, H., Schröder, E., Langreth, D. C. & Lundqvist, B. I. Van der Waals Density Functional for General Geometries. In: *Physical Review Letters* 92.24 (2004), p. 246401. URL: <https://doi.org/10.1103/PhysRevLett.92.246401>.
- [56] Batatia, I., Kovács, D. P., Simm, G. N. C., Ortner, C. & Csányi, G. “MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields”. In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 11423–11436. arXiv: [2206.07697](https://arxiv.org/abs/2206.07697). URL: <https://arxiv.org/abs/2206.07697>.
- [57] Zhou, W., Zheng, D., Liu, Q., Lu, D., Liu, Y., Lin, P., Huang, Y., Peng, X., Bao, J. J., Cai, C., Jin, Z., Wu, J., Zhang, H., Jin, G., Ji, Y., Shen, Z., Liu, X., Sun, L., Cao, Y., Sun, M., Liu, J., Chen, T., Liu, R., Li, Y., Han, H., Liang, X., Bao, T., Deng, Z., Liu, T., Chen, N., Ren, H., Zhang, X., Liu, Z., Fu, Y., Liu, M., Li, Z., Wen, T., Tang, Z., Xu, Y., Duan, W., Wang, X., Gu, Q., Dai, F.-Z., Zheng, Q., Zhong, Y., Xiang, H., Gong, X., Zhao, J., Zhang, Y., Ou, Q., Jiang, H., Liu, S., Xu, B., Xu, S., Ren, X., He, L., Zhang, L. & Chen, M. ABACUS: An electronic structure analysis package for the AI era. In: *The Journal of Chemical Physics* 163.19 (2025), p. 192501. URL: <https://doi.org/10.1063/5.0297563>.
- [58] Ong, S. P., Richards, W. D., Jain, A., Hautier, G., Kocher, M., Cholia, S., Gunter, D., Chevrier, V. L., Persson, K. A. & Ceder, G. Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. In: *Computational Materials Science* 68 (2013), pp. 314–319. URL: <https://doi.org/10.1016/j.commatsci.2012.10.028>.