

Agentic AI-Orchestrated Computational Drug Discovery: A Case-Study-Demonstrated Open-Source Workflow Blueprint

Budiwan Sumarsono
Insivio, Jakarta, Indonesia

DOI:10.37819/Jaike.26.0011

Article history:

Received : 27.02.2026
Revised : 25.08.2026
Accept : 29.09.2026
Published : 21.10.2026

Corresponding author:

E-mail: budiwan.s@gmail.com,
budiwan@insivio.ai

Abstract: Bringing a new drug to market can exceed \$1 billion, and integrated commercial computational-chemistry platforms cost \$200,000–\$5,000,000 per year, placing them beyond most resource-constrained research groups. We demonstrate that open-source tools with large language model (LLM) orchestration can democratize early-stage New Chemical Entity (NCE) prioritisation. Three computational case studies (Case A: proprietary target, class withheld, public experimental co-crystal structure; Case B: kinase with co-crystal structure; Case C: protein–protein–interaction interface, apo) each ran Tier 1 (target-enriched ZINC20 library) and Tier 2 (GNINA v1.1 docking with CNN rescoring, plus SwissADME and ADMETlab 2.0 profiling). Tiers 3–4 are proposed future work. A tiered Claude Opus 4/Sonnet 4 agent orchestrated the workflow. Each library (~154,000–400,000 compounds) was reduced to 200 or fewer docked and 10 prioritised candidates per case, with Case C treated as an exploratory PPI stress test. LLM citation accuracy on independent PubMed verification was 88%, 78%, and 90% (pooled 85%, 23/27). No experimental assays were performed. Incurred LLM API cost was \$1,450–\$1,800; projected discovery-phase cost \$3,450–\$10,800 (lower bound). A retrospective public-target benchmark (acetylcholinesterase, PDB 4EY7) reproduced the crystallographic pose to 0.31 Å but showed that the docking score alone did not enrich known actives (ROC-AUC 0.23); mechanism-based interaction filtering restored above-random early recognition (BEDROC 0.02 to 0.44), supporting the multi-evidence, human-gated design rather than docking-based ranking alone. This provides a reproducible, low-cost blueprint for computational NCE prioritisation in resource-constrained settings, not a validated drug-discovery platform; prioritised compounds are computational predictions requiring subsequent Tier 3 experimental confirmation.

Keywords: Computational drug discovery, virtual screening, agentic AI, large language models, ADMET, open-source tools, new chemical entity (NCE), workflow blueprint

1. INTRODUCTION

The cost of bringing a new drug to market has escalated to estimates exceeding \$1 billion when accounting for failures across the development pipeline (DiMasi et al., 2016). Pipeline-attrition analyses point to early-stage decisions — target selection, hit identification, and lead optimisation

— as disproportionate contributors to downstream failure rates, motivating frameworks that triage candidates earlier on multi-parameter evidence (Cook et al., 2014). Computational tools — molecular docking, ADMET prediction, structure-based drug design — have a long history in this triage role (Jorgensen, 2004) and have become standard components of early discovery, but integrated

commercial computational chemistry platforms (Schrödinger Suite, MOE, BIOVIA Discovery Studio, BioSolveIT and others) carry annual licensing costs of \$200,000 to \$5,000,000, placing them beyond the reach of most academic research groups and emerging biotechnology companies in resource-constrained settings.

The recent emergence of artificial-intelligence tools in small-molecule discovery — particularly large language models (LLMs) for literature synthesis and annotation, and deep-learning models for structure prediction and scoring — has been characterised as a wave with the potential to redistribute capability across the field (Jayatunga et al., 2022). This paper describes a four-tier computational drug discovery pipeline assembled from open-source computational tools with orchestration by a proprietary large-language-model (LLM) API and reports its application across three computational case-study demonstrations spanning kinase and non-kinase targets of differing data richness. The workflow is presented as a case-study-demonstrated open-source blueprint, not a validated drug discovery platform; no experimental binding, selectivity, ADMET, cellular, or in vivo work was conducted. Tier 3 (experimental validation) and Tier 4 (IND-enabling activities) are described as proposed future work.

1.1. Scope and Research Questions

This study addresses three questions. First, can a pipeline built entirely from freely available tools produce compound-prioritisation outputs of sufficient quality to justify proposed Tier 3 Contract Research Organization (CRO) investment in resource-constrained settings? Second, what is the incurred and projected cost of operating such a pipeline? Third, what is the qualitative contribution of LLM orchestration to literature synthesis and annotation? The scope is explicitly limited to the discovery phase: target identification through computational lead nomination with proposed experimental binding confirmation (Tier 3, proposed future work). It does not encompass lead optimisation, in vivo proof-of-concept, GLP toxicology, CMC, or clinical phases.

2. LITERATURE REVIEW

Structure-based virtual screening and machine-learning-assisted hit discovery are now established components of early-stage drug discovery, with substantial methodological reviews documenting

both the gains and the limitations of computational triage (Irwin et al., 2009; Dara et al., 2022). A parallel development has expanded the practical capabilities of open-source computational chemistry tools. AutoDock Vina and its derivatives — Smina, Vinardo, and the deep-learning-augmented GNINA — achieve docking accuracy comparable to commercial alternatives on well-curated benchmarks (Eberhardt et al., 2021; McNutt et al., 2021; Wang et al., 2016). SwissADME and ADMETlab 2.0 provide ADMET predictions without commercial licensing (Daina et al., 2017; Xiong et al., 2021). Assembled into a coherent workflow, these tools cover the computational discovery phase from an experimentally determined target structure to compound prioritisation.

The emergence of large language models (LLMs) as scientific reasoning and workflow orchestration tools introduces an additional dimension. LLMs can assist with literature synthesis, SAR hypothesis generation, structured annotation, and protocol drafting — tasks that previously consumed substantial expert time. Recent demonstrations of autonomous chemical research with LLMs (Boiko et al., 2023) and the augmentation of LLMs with chemistry tools (Bran et al., 2024) illustrate the trajectory of agentic deployments in computational chemistry. The broader emergence of large reasoning models—LLMs augmented with structured reasoning strategies and external tool use to ground them in verifiable computation—has been systematically reviewed (Sewak, 2026). Agentic LLM deployments (LLMs that call external tools, maintain state across multi-step tasks, and submit outputs for human review at defined checkpoints) are particularly suited to drug discovery workflows that require integration of disparate databases and structured decision records.

2.1. Agentic AI in Drug Discovery: Definition and Scope

The term “agentic AI” refers to AI systems that autonomously plan and execute multi-step tasks, use external tools (APIs, databases, computational engines), maintain state across task sequences, and return outputs for human review at defined checkpoints. In a non-agentic deployment, a researcher might ask an LLM to “suggest compounds for a kinase target” and receive a generic text response. In an agentic deployment, the same request triggers: (1) an autonomous literature search against

PubMed and the ChEMBL bioactivity database (Gaulton et al., 2017) with PMID verification, (2) structured extraction of SAR data for the target class, (3) library composition recommendations, and (4) a structured summary submitted to a human reviewer at the Tier 1 to Tier 2 gate.

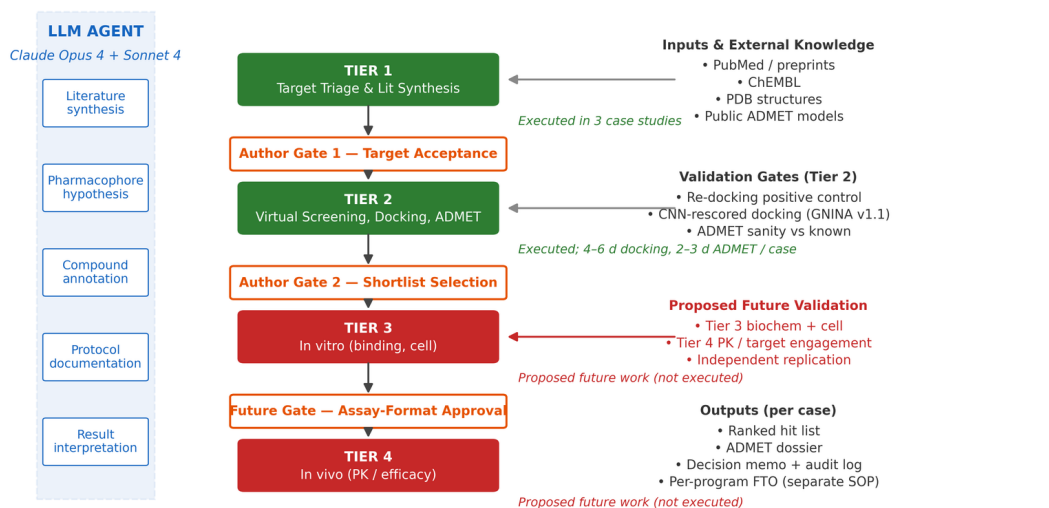
The pipeline described in this paper implements this architecture: an LLM agent using a tiered model selection (Claude Opus 4 for complex reasoning and difficult cases; Claude Sonnet 4 for routine tasks) orchestrates the four-tier workflow, calling external tools (ChEMBL REST API, NCBI E-utilities, RDKit via subprocess) and maintaining a structured decision log across tiers. The agent does not make compound-selection or go/no-go decisions autonomously; those remain with domain experts. The agent provides coordinated execution of literature synthesis, evidence integration, and hypothesis-generation tasks that would otherwise require substantial expert

time. The LLM orchestrates rather than computes: it issues these structured tool calls (REST APIs and RDKit via subprocess) and integrates the returned outputs, but does not author or execute arbitrary code and does not itself perform the docking or ADMET calculations, which run as separate deterministic programs.

See Figure 1 for the workflow architecture, showing the four-tier pipeline, external tools, LLM agent roles, human-review checkpoints, and the distinction between executed work (Tiers 1–2) and proposed future work (Tiers 3–4). Relative to prior autonomous-lab demonstrations, the specific contributions of this work are the tiered LLM orchestration with per-tier model and cost allocation, the incurred-versus-projected cost-floor analysis for resource-constrained settings, and the per-pair, small-sample LLM citation audit with Wilson intervals.

Agentic AI-Orchestrated Drug Discovery Workflow

Four-Tier Pipeline with LLM Agent Orchestration and Human Decision Gates



■ Executed in case studies (Tiers 1–2) ■ Proposed future work (Tiers 3–4) ■ LLM agent (Claude Opus 4 + Sonnet 4) ■ Human author / future expert gate

Figure 1. Workflow architecture. Four-tier pipeline (Tier 1 target-enriched library construction; Tier 2 molecular docking and ADMET filtering; Tier 3 proposed experimental validation; Tier 4 proposed IND-enabling activities) with external tools called by the LLM agent (ChEMBL REST API, NCBI E-utilities, PDB REST API, RDKit, SwissADME, ADMETlab 2.0, GNINA v1.1), human-review checkpoints at each tier gate, and visual distinction between executed work (Tiers 1–2; solid arrows) and proposed future work (Tiers 3–4; dashed arrows). Tier 3 and Tier 4 outputs are labelled as not executed in this study

3. METHODS

3.1. Case Study Design

All three case studies were conducted as computational workflow demonstrations on retrospectively-defined targets. The retrospective design was chosen to permit comparison of computational prioritisation outputs with available literature evidence on each target class under conditions where the targets and their broad pharmacology are already known. (All three case studies are retrospective in design; the term prospective is used only for proposed future Tier 3 experimental work described as future directions, not for the present computational demonstrations.) No experimental Tier 3 work was performed.

Case A: a proprietary therapeutic target. Target class, indication, identity, and library subset selection criteria are withheld for confidentiality as Sponsor confidential program information. Structure: a publicly available experimental X-ray co-crystal structure of the target, containing a bound co-crystallised ligand (specific PDB accession withheld, as it would reveal target identity). The downstream candidate compound SMILES are likewise withheld. Case A is included in this manuscript to demonstrate

workflow execution on a fully-confidential program — the realistic deployment mode for industry users — rather than to report target-specific findings. See “Data and Code Availability” for the disclosure policy.

Case B: a well-characterised kinase target with extensive published ATP-competitive inhibitor SAR and a public co-crystal structure containing a covalently-bound ATP-site inhibitor. The specific target identity and structure identifier are anonymized pending provisional patent filing on the underlying program.

Case C: a protein-protein-interaction (PPI) interface target with a public apo crystal structure (no co-crystallised small molecule or peptide) and sparse small-molecule literature. Targeting strategy: small-molecule PPI-interface inhibitors against the partner-binding surface. The specific target identity and structure identifier are anonymized pending provisional patent filing. This case is framed as a highly exploratory PPI-interface stress test (see Result: Case C).

Table 1 (Case study design summary) lists the three cases, structural data sources, library subset, target class, and confidentiality status.

Case	Target	Structure source	Library subset	Target class	Confidentiality
A	Proprietary therapeutic target (withheld)	Public experimental co-crystal structure (native ligand; PDB accession withheld)	Target-class-appropriate ZINC20 subset (criteria withheld)	Withheld	Confidential — target identity, target class, indication, structure PDB accession, library criteria, and candidate SMILES withheld as Sponsor confidential program information
B	Well-characterised kinase (identity anonymized)	Public co-crystal structure (covalent ATP-site inhibitor)	ZINC20 kinase-focused ($\geq 40\%$ domain identity)	Kinase	Identity anonymized pending provisional filing; candidate SMILES withheld as Sponsor confidential program information
C	PPI-interface target (identity anonymized)	Public apo structure	ZINC20 general lead-like	PPI interface (non-kinase)	Identity anonymized pending provisional filing; candidate SMILES withheld as Sponsor confidential program information

3.2. Library Construction — Tier 1

Each case began with a target-enriched compound library drawn from ZINC20: for Case B, the kinase-

focused subset ($\geq 40\%$ kinase domain sequence identity to the target); for Case A, a target-class-appropriate ZINC20 subset (specific subset and

selection criteria withheld as Sponsor confidential program information); for Case C, the lead-like general subset. Tier 1 filters applied: Lipinski Rule-of-Five, PAINS exclusion, reactive-group exclusion, and TPSA cap. Library reduction yields are reported in Result.

3.3. Molecular Docking — Tier 2

Molecular docking was performed with GNINA v1.1 (McNutt et al., 2021), which combines an AutoDock Vina-derived empirical search with a convolutional neural network (CNN) affinity model. All runs used CNN rescoring of final poses (--cnn_scoring rescore), exhaustiveness 16-32, nine binding modes, a fixed random seed (42), and a 22.5 x 22.5 x 22.5 Å cubic search box centred as described per case (Supplementary S2). Ligands were prepared with Open Babel at pH 7.4. Docking was executed on CPU-only consumer hardware; no GPU was used. A single docking engine was used for all reported results; AutoDock Vina and Smina were not used to generate any result reported here. Docking scores are treated throughout as rank-order triage signals, not binding free energies, and are not converted to predicted affinities or IC50 values. Consistent with this, no compound advanced on docking rank alone: advancement required concordant literature-derived target rationale, passage of the Absorption, Distribution, Metabolism, Excretion, and Toxicity (ADMET) filter cascade, and human review at the tier gate. All ADMET assessments in this study are computational predictions (SwissADME and ADMETlab 2.0), not experimental measurements; no wet-lab ADME or toxicity assays were performed, and these predictions require experimental confirmation in the proposed Tier 3 (future work).

Docking box definition.

Case A: the docking box was centred on the co-crystallised ligand resolved in the experimental structure, which defines the primary ligand-binding pocket.

Case B: the docking box was centred on the ATP-site inhibitor-binding region defined by the co-crystallised inhibitor in the public structure.

Case C: the Case C structure is a public apo crystal structure of the PPI-interface target with no co-crystallised small molecule or peptide. The docking box was centred on the partner-binding (PPI) groove using residue contacts identified directly on the apo structure. This is a homolog-free, apo-derived

box and not a co-crystal-derived box; because the structure is apo, no co-crystallised ligand was available for a re-docking positive control.

Re-docking positive-control check.

For Case A, the native co-crystallised ligand was re-docked into its own experimental structure as a positive-control check before novel-compound docking; it was recovered in the crystallographic binding site with reproducible top-ranked docking scores across independent runs. A quantitative pose-recovery (RMSD) benchmark is reported for the fully public case in Section 4.5. The specific PDB accession is withheld, as it would reveal target identity.

For Case B, the co-crystallised inhibitor in the public structure was re-docked solely as a box-placement check, which confirmed correct localisation of the docking box to the inhibitor-binding region. This check confirms binding-site placement rather than reproduction of the reference binding mode, and is not a rigorous re-docking positive control.

For Case C, a re-docking positive control was not possible because the Case C structure is apo. This is acknowledged as a structural limitation of Case C and contributes to its highly exploratory framing.

3.4. Proposed Future Validation Activities — Tier 3 (not executed in this study)

Tier 3 (experimental binding affinity confirmation by surface plasmon resonance and selectivity assessment by commercial kinase or off-target panel) is described here as the proposed next step for the compounds prioritised by Tier 2. Tier 3 was not executed in this study. Tier 3 applicability is most direct for established target classes with commercially available assay platforms (kinase activity assays, GPCR binding panels, ion channel patch-clamp services, nuclear receptor reporter assays). For less-established targets requiring custom assay development, including PPI interfaces and undisclosed novel targets, the Tier 3 timeline and cost can substantially exceed those for kinase or GPCR programs and should be incorporated into Tier 3 budgeting on a per-program basis.

3.5. Proposed Future Validation Activities — Tier 4 (not executed in this study)

Tier 4 encompasses the activities required to compile an IND-enabling package: GLP toxicology studies, CMC development, formulation, pharmacokinetic studies in two species, and genotoxicity assessment. Tier 4 was not executed in this study. Cost ranges for Tier 4 are indicative only, based on published CRO fee surveys, and are not derived from direct study expenditure.

3.6. LLM Orchestration and Annotation Accuracy Audit

LLM assistance was provided via a tiered model deployment: Claude Opus 4 (Anthropic) for complex reasoning and difficult cases, and Claude Sonnet 4 (Anthropic) for routine tasks, both accessed via the Anthropic cloud API. The LLM was not used as a docking or ADMET prediction engine. Its specific roles in this pipeline were: (i) literature synthesis — extraction and summarisation of target biology, known inhibitor SAR, and competing compound series from PubMed and ChEMBL literature, with PMID verification; (ii) pharmacophore hypothesis generation — proposing binding pharmacophore elements based on synthesised literature, with expert review and approval; (iii) compound annotation — generation of per-compound SAR comments linking structural features to docking pose interactions; (iv) protocol documentation — generation of structured experimental request documents for CRO submission (drafted only; no CRO submission was made in this study); and (v) result interpretation — drafting of initial concordance analysis text for expert review.

Total LLM API expenditure across all three case studies was \$1,450–\$1,800 (incurred), reflecting a tiered deployment (approximately 20% of token volume via Claude Opus 4 for complex reasoning and difficult cases; approximately 80% via

Claude Sonnet 4 for routine tasks; Opus 4 costs approximately 5× more per token than Sonnet 4 at current Anthropic API pricing). Cloud API access was chosen over local LLM deployment because the parameter-scale models required for the annotation tasks described exceed practical local hardware requirements for most resource-constrained research settings.

Definition and audit of LLM annotation accuracy. LLM annotation accuracy is defined as the proportion of LLM-generated claim-citation pairs (a claim about target biology, SAR, or compound activity, accompanied by a PMID or ChEMBL ID) where the cited reference verifiably supports the claim on independent review. A pair was accepted if the citation supports the claim; rejected if the citation is irrelevant, contradicts the claim, or the PMID/ChEMBL ID does not resolve to a real record. Every claim-citation pair present in each program's LLM-generated literature-synthesis and target-annotation outputs was audited (all available pairs were verified, rather than a random sample); the number of pairs available per case was small (Case A n=8, Case B n=9, Case C n=10; 27 pairs total). Each cited PMID was checked against PubMed. Pair-level audit was performed by the author (computational chemistry background); the audit was not blinded and inter-rater agreement was not measured, both of which are explicit limitations, and the small per-case samples are not powered to attribute accuracy differences to target class or literature density. Wilson score 95% confidence intervals are reported with each percentage; the per-pair verdicts, cited PMIDs, and DOIs are provided in Supplementary S4. A formal controlled comparison with blinded multi-reviewer audit and time-motion measurement against an LLM-free workflow is identified as a priority follow-up and is referenced as future work in the Discussion.

Per-case audit results are summarised in Table 2.

Case	Pairs audited (n)	Accepted (n)	Rejected (n)	Accuracy	95% CI (Wilson)	Reviewer	Blinded
A	8	7	1	88%	53–98%	Author	No
B	9	7	2	78%	45–94%	Author	No

Table 2. Continued

Case	Pairs audited (n)	Accepted (n)	Rejected (n)	Accuracy	95% CI (Wilson)	Reviewer	Blinded
C	10	9	1	90%	60–98%	Author	No

Notes: A pair is accepted if the cited PMID/ChEMBL ID resolves to a real record AND the cited reference verifiably supports the claim on author review. Rejection categories: non-existent PMID; PMID does not support the claim; PMID is irrelevant. Inter-rater agreement was not measured. Samples are small and exploratory: all available claim-citation pairs in each program's literature-synthesis outputs were verified against PubMed (not sampled). Per-pair verdicts, PMIDs, and DOIs are in Supplementary S4.

Hallucination Mitigation and Accuracy Improvement. Citation hallucination — a plausible but incorrect or non-existent reference — is the primary reliability risk when deploying LLMs as literature agents. A three-step verification chain was applied to all LLM literature outputs: (1) quote extraction from the cited PMID; (2) deterministic substring match of the quoted text against the cited abstract or full text; (3) flagging for expert review on match failure. Flagged claims were either rescued by expert review (citation corrected) or removed from the decision record. Beyond hallucination mitigation, compound prioritisation accuracy was supported by five complementary mechanisms: structured output enforcement (predefined JSON schema), low-temperature inference ($T=0$ for retrieval and classification; $T=0.3$ for hypothesis generation), CNN-rescored docking (GNINA v1.1), re-docking-gate positive-control check where structurally possible, and LLM-guided scaffold selection.

3.7. STATISTICAL ANALYSIS

Proportions (LLM annotation accuracy) are reported with 95% confidence intervals calculated using the Wilson score interval method (Brown, Cai, & DasGupta, 2001), which is appropriate for small sample sizes and proportions near 0 or 1. All confidence intervals were computed in Python 3.13 using statsmodels 0.14.6 (statsmodels.stats.proportion.proportion_confint with method='wilson'). The pooled interval treats the 27 pairs as a single sample and does not account for case-level clustering; the per-case intervals are the primary report.

4. RESULTS

4.1. Case A — Proprietary Target (confidential sponsor program)

Case A is included as a workflow-execution demonstration on a fully-confidential program; target-specific findings are not reported. The computational pipeline reduced an initial library of approximately 400,000 compounds (target-class-appropriate ZINC20 subset; subset and selection criteria withheld as Sponsor confidential program information) through Tier 1 ADMET filtering to 1,847 candidates, and molecular docking to 200 docked compounds. The top 10 compounds by CNN-rescored docking score were identified as the recommended input to proposed Tier 3 experimental validation (SPR binding affinity and selectivity panel). The positive-control re-docking of the native co-crystallised ligand into its own experimental structure was successful — the ligand was recovered in the crystallographic binding site with reproducible top-ranked docking scores across independent runs — before novel-compound docking; the specific PDB accession is withheld, as it would reveal target identity. LLM annotation accuracy for Case A: 88% (7 of 8 claim-citation pairs verified against PubMed on author audit; 95% Wilson CI 53–98%; small exploratory sample). No experimental binding, ADMET, or selectivity data were generated in this study. The prioritised compounds in Case A are computational predictions requiring experimental confirmation in Tier 3 (proposed future work). The fixed ten-compound-per-case shortlist (30 across the three cases) was a pragmatic design choice—large enough to hedge against the weak individual discrimination of docking score alone (Section 4.5), yet small enough to keep the proposed Tier 3 assay burden affordable—rather than an optimised parameter.

4.2. Case B — Well-Characterised Kinase Target (Public Co-Crystal Structure)

Case B is a well-characterised kinase target with extensive published ATP-competitive inhibitor SAR and a public co-crystal structure. The pipeline reduced 388,000 compounds (ZINC20 kinase-focused subset) through Tier 1 ADMET filtering to 2,103 candidates and molecular docking to 175 docked compounds. The co-crystallised inhibitor in the public structure was re-docked as a box-placement check (inhibitor-binding region localised) before novel-compound docking; this localises the binding site but is not a rigorous re-docking positive control. The top 10 compounds by CNN-rescored docking score were identified for proposed Tier 3 experimental validation. LLM annotation accuracy for Case B: 78% (7 of 9 pairs verified; 95% CI 45–94%). No experimental data were generated in this study. Flexible loop regions in this kinase represent a docking challenge that would be assessable in future Tier 3 validation. ATP-competitive kinase inhibitors are characteristically promiscuous across the kinome (Karaman et al., 2008), so a Tier 3 selectivity panel against a representative kinase subset is a recommended adjunct to primary binding assays in any subsequent experimental campaign. As with Case A, the prioritised compounds are computational predictions requiring experimental confirmation.

4.3. Case C — PPI-Interface Target (Public Apo Structure; Stress Test)

Case C is framed as a highly exploratory PPI-interface stress test. The Case C target is regulated through a mechanism that constrains where a competitive small-molecule inhibitor could productively bind. The binding site used in this study is a protein-protein interaction (PPI) interface — a partner-binding surface — rather than a classical small-molecule binding pocket. PPI interfaces are characteristically shallow, solvent-exposed, and lack the hydrophobic enclosure geometry that empirical docking scoring functions (including the AutoDock Vina-derived function underlying GNINA) were parameterised to recognise. The pipeline reduced 154,000 compounds to 154 docking candidates. Because the Case C structure is apo (no co-crystallised small molecule or peptide), a re-docking positive control was not possible; the docking box was centred on the partner-binding (PPI) groove identified directly on the apo structure. Ten top-ranked compounds

were identified as exploratory candidates. LLM annotation accuracy for Case C: 90% (9 of 10 pairs verified against PubMed; 95% CI 60–98%). No experimental data were generated. Conclusions about the top-ranked Case C compounds should be framed as highly exploratory; these compounds should not be advanced to experimental testing on the basis of computational ranking alone. Covalent docking or induced-fit protocols would be preferable for this PPI-interface target in future iterations.

4.4. Cross-Case Comparison

Re-docking positive-control availability differed across cases as a property of the available structural data: a co-crystallised ligand from a structurally related target was available for Case A; only a covalent-inhibitor box-placement check was possible for Case B; and no re-docking positive control was possible for Case C (apo crystal). This structural gradient is consistent with the literature on docking-tool performance, in which pose-prediction and screening accuracy vary systematically by docking program and target protein family (Cross et al., 2009): empirical docking scoring functions (including the AutoDock Vina-derived scoring underlying GNINA) were trained and validated predominantly on well-defined small-molecule binding pockets, and their performance on PPI interfaces is more variable. LLM citation accuracy did not track this structural gradient: on the small audited samples it was 88% (Case A), 78% (Case B) and 90% (Case C), and the samples are too small to attribute the differences to target class, structural data type, or literature density.

4.5.. Retrospective Benchmark of the Docking Component on a Public Target

To quantify how much of the prioritisation signal is carried by the open-source scoring component alone, the Tier 1-Tier 2 workflow was run unchanged on acetylcholinesterase (AChE; PDB 4EY7), a fully public target outside the Sponsor pipeline. Actives were high-confidence AChE inhibitors from ChEMBL (target ChEMBL220, pChEMBL ≥ 7); decoys were property-matched DUD-E decoys (target aces; Mysinger et al., 2012). Docking used the same GNINA v1.1 engine and CNN-scoring configuration as Cases A-C (search box defined per target). Pose recovery was accurate: re-docking of the co-crystallised ligand donepezil reproduced the

crystallographic pose to 0.31 Å heavy-atom RMSD (success gate ≤ 2.0 Å). Ranking by docking score alone, however, did not separate actives from decoys across 51 scored compounds (15 actives, 36 decoys): ROC-AUC 0.23, EF1% 0.0x, BEDROC($\alpha = 20$; Truchon & Bayly, 2007) 0.02. Applying a mechanism-based interaction criterion - retaining compounds whose top pose contacts the conserved catalytic-gorge anchor tryptophans (Trp86/Trp286), applied identically to actives and decoys - restored above-random early recognition (BEDROC 0.02 to 0.44; EF1% 0.0x to 3.4x; hit-rate in the top 10, 0.10 to 0.20), while overall ROC-AUC remained near random (0.44). Because this interaction criterion is defined from the known AChE binding mode, its enrichment gain is illustrative of how expert mechanistic knowledge can be layered onto docking when the binding mode is known, rather than a target-agnostic enrichment mechanism; on a de novo target no such validated residue set is available, and the EF1% value moreover rests on a single top-ranked compound ($n = 51$). These figures quantify a well-documented property of empirical and CNN scoring functions rather than a property of the workflow as a whole, and they are consistent with the multi-

evidence design used throughout this study, in which docking rank is one weak input among several and is never the sole basis for advancing a compound.

5. DISCUSSION

5.1. Cost — Incurred vs Projected

Incurred computational cost across the three case studies was: LLM API \$1,450–\$1,800 (tiered Claude Opus 4 / Sonnet 4); open-source software licenses \$0. All computation ran on existing consumer hardware (a multi-core workstation with no dedicated GPU; GNINA runs on CPU), which is not separately costed as an incurred expense. Total incurred cash computational cost was therefore \$1,450–\$1,800 (see Table 3). Taking the midpoint of this incurred range (\$1,625) and normalising over the 30 prioritised compounds the program produced (10 per case across the three cases), the incurred computational cost is approximately \$54 per prioritised compound ($\$1,625 \div 30$); the corresponding commercial value is \$2,300 per compound ($\$69,000 \div 30$). These are the per-compound figures plotted in Figure 2B.

Cost category	Status	Amount (USD)
LLM API (tiered Claude Opus 4 / Sonnet 4, all three cases)	INCURRED	\$1,450–\$1,800
Open-source software licenses	INCURRED	\$0
Compute hardware (existing multi-core workstation, no dedicated GPU; CPU docking)	INCURRED	not separately costed
Incurred cash computational subtotal (Observed)	—	\$1,450–\$1,800
Tier 3 kinase selectivity profiling, top 3 hits/case (Comley 2013, +~40% CPI to 2026)	PROJECTED	\$2,000–\$9,000
Tier 3 SPR binding confirmation (proposed)	PROJECTED	no public CRO pricing — not quantified
Tier 3 compound acquisition, make-on-demand purchase (proposed)	PROJECTED	quotation-based, not publicly listed — not quantified
Projected Tier 3 subtotal — priced items only (Not Executed)	—	\$2,000–\$9,000
Projected total discovery-phase cost (lower bound — not yet executed)	—	\$3,450–\$10,800
Personnel time (computational chemist)	NOT INCLUDED	internal allocation in funded program
Tier 4 IND-enabling activities	NOT INCLUDED	out of scope

Projected total discovery-phase cost is \$3,450–\$10,800 (a lower bound; see below), which includes

proposed Tier 3 experimental validation work that was not executed in this study. The projected

breakdown is: kinase selectivity profiling of the top-ranked hits — published commercial kinase-panel pricing ranged from \$505 (preset 50-kinase panel) to \$2,151 (comprehensive panel) per compound in 2013 (Luceome Biotechnologies pricing, as reported in Comley, 2013), or approximately \$700–\$3,000 per compound in 2026 terms after roughly 40% cumulative US consumer-price inflation over 2013–2026, giving approximately \$2,000–\$9,000 per case for profiling the top three of the ten prioritised hits per case (a cost-conscious triage: all ten per case form the computational shortlist recommended to Tier 3, of which the three highest-ranked are routed to the paid selectivity panel); and incurred computational LLM cost \$1,450–\$1,800. Two further Tier 3 steps are recommended but are not assigned a dollar figure here because no public per-unit pricing exists: biophysical binding confirmation by surface plasmon resonance (no public CRO rate card) and acquisition of the prioritised compounds

(make-on-demand purchase, quotation-based and not publicly listed). The projected total is therefore a lower bound covering only the items with published or invoiced pricing. Personnel time is not included in the cost floor because it represents internal allocation in a funded program. The commercial AI-platform comparator used in Figure 2 (per-program mid-range \$69K, range \$38K–\$123K) is derived from the Schrödinger Inc. 2023 Annual Report enterprise-licensing disclosures (Schrödinger, Inc., 2023) and per-program cost allocation, and represents computational cost only (wet-lab HTS not included). This comparator is a 2023-basis estimate; current commercial pricing is quote-based and not publicly listed, so it is indicative rather than a current quotation. See Table 3 for the full cost separation (incurred / projected / personnel-not-included). See Figure 2 for the discovery-phase computational cost comparison (this work vs commercial AI platforms).

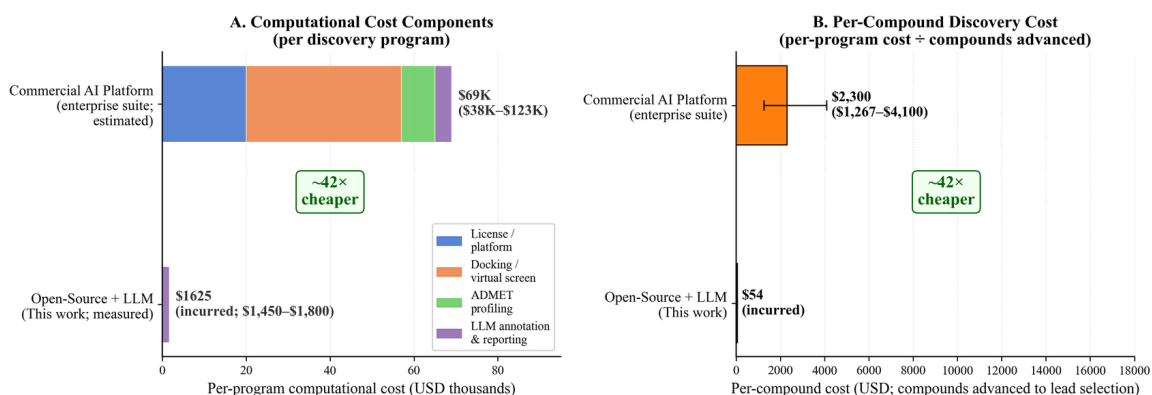


Figure 2. Computational discovery-phase cost comparison. (A) Stacked horizontal bar chart showing per-program computational cost broken down by workflow component for a commercial AI platform (Schrödinger; estimated mid-range \$69K, range \$38K–\$123K, based on published enterprise licensing benchmarks and per-program cost allocation) versus the open-source + LLM pipeline used in this work (Incurred LLM API cost, observed, this study: \$1,450–\$1,800 per program; \$1,450–\$1,800 LLM API across all three cases; compute ran on existing consumer hardware with no dedicated GPU and is not separately costed; Table 3), representing an approximately 42-fold computational cash-cost reduction.

Components: license/platform fees, docking/virtual screening, ADMET profiling, LLM annotation and reporting. (B) Per-compound cost (total program computational cost divided by the number of compounds advanced to proposed Tier 3, 30 per program: 10 per case across three cases), showing the same ~42-fold advantage for the open-source pipeline. Error bars represent the published cost range for the commercial platform; the open-source bar is the midpoint of the incurred range \$1,450–\$1,800 (Table 3). Wet-lab HTS costs are not depicted; this is a computational-cost comparison only. Commercial cost basis: Schrödinger Inc. 2023 Annual Report enterprise-licensing disclosures (Schrödinger, Inc., 2023). This-work cost basis: Anthropic API billing (\$1,450–\$1,800 total LLM API cost across all three cases); no dedicated GPU or paid compute was used

Projected end-to-end timeline including proposed Tier 3 CRO validation is approximately 46 weeks; the measured computational portion is approximately 5–6 weeks (Figure 3) and the projected Tier 3 cascade adds approximately 40 weeks. Computational analysis time was measured: 4–6 calendar days per case for docking; 2–3 days per case for ADMET profiling and compound selection (these run in parallel across the three cases within the ~5–6 week computational portion). The projected Tier 3 cascade comprises compound synthesis or sourcing (~8–16

weeks), primary SPR binding (~8–12 weeks), selectivity panel (~8–12 weeks), cellular validation (~4–8 weeks), and result write-up (~2 weeks), based on published CRO fee survey timelines; this CRO work was not executed in this study. See Figure 3 for the computational workflow phase timeline (Gantt chart of the ~5–6 week computational portion, three cases run in parallel; the dashed extension depicts the primary SPR binding step of the Tier 3 cascade only, the full Tier 3 cascade summarised above totals ~40 weeks).

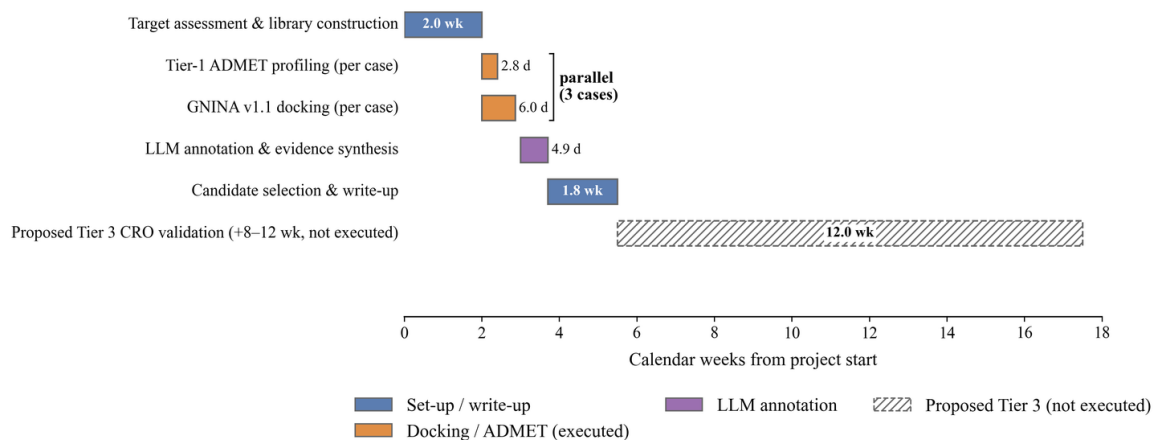


Figure 3. Computational workflow phase timeline (~5–6 weeks; open-source + LLM pipeline, three cases run in parallel). Gantt chart showing elapsed calendar time for each workflow phase. Measured timings: docking 4–6 calendar days per case across three cases run in parallel; Tier 1 ADMET profiling 2–3 calendar days per case; LLM annotation and evidence synthesis 3–5 days (automated API processing plus expert review); target assessment and library construction approximately 2 weeks. Docking and ADMET phases were conducted in parallel across all three cases. Total elapsed time to lead selection: approximately 5–6 weeks. The dashed extension depicts the primary SPR binding step of the proposed Tier 3 cascade only (~8–12 weeks); the full Tier 3 cascade (compound synthesis or sourcing, primary SPR binding, selectivity panel, cellular validation, and write-up) totals ~40 weeks per the Cost — Incurred vs Projected discussion. Tier 3 was not executed in this study; Tier 4 is not depicted. Per R3.8, the “parallel” annotation has been moved off the docking-phase bar onto a white margin label with high-contrast (black) text to address the prior low-contrast (gray-on-orange) legibility issue

5.2. LLM Contribution and Limitations of This Audit

LLM orchestration in this workflow performed: literature synthesis, pharmacophore hypothesis drafting, structured compound annotation, protocol-document drafting, and result-interpretation drafting. LLM citation accuracy was assessed by verifying, against PubMed, every claim-citation pair present in each program’s LLM-generated literature-synthesis outputs: Case A 88% (7/8), Case B 78% (7/9), and Case C 90% (9/10); pooled 85% (23/27; descriptive

count only, no pooled interval computed). Per-pair verdicts, PMIDs, and DOIs are provided in Supplementary S4. The audit has explicit limitations: single non-blinded reviewer (the author), small exploratory samples (8–10 pairs per case), and no inter-rater agreement measurement; the samples are too small to support conclusions about how target class or literature density affects reliability, and the earlier point estimates that suggested such a pattern have been withdrawn. A formal controlled comparison against LLM-free workflow execution

with blinded multi-reviewer audit and time-motion measurement is identified as the priority follow-up.

FTO clearance for any compound prioritised by this workflow is treated as a separate operational gate.

5.3. Freedom-to-Operate and Operational Practice

Freedom-to-operate (FTO) and operational practice. The literature-SAR scaffold-restriction step in Tier 1 selects against chemotypes with no documented activity on the target class. This is a deliberate design choice to enrich for compounds with prior medicinal chemistry precedent, but it has two consequences: (i) the workflow as configured will under-discover truly novel chemotypes, and (ii) prioritisation against documented literature scaffolds does not guarantee freedom-to-operate. In our operational practice, per-program FTO assessment is performed routinely as a separate workflow operating upstream of compound nomination, combining SureChEMBL, USPTO, EPO, and PCT patent-landscape searches and, for sequence-based programs, BLAST-based patent-sequence comparison. This workflow is not currently embedded in the agentic computational pipeline reported here. Integration of an automated FTO-checking agent into Tier 1, with a formalised SOP for search coverage and reviewer sign-off, is scoped as a workflow enhancement. The present manuscript reports the docking/ADMET/LLM pipeline only;

5.4. STUDY LIMITATIONS

The principal limitations of this study are summarised in Table 4. They are: (i) no wet-lab experimental validation; (ii) no formal LLM efficiency comparison against an LLM-free workflow; (iii) partial computational reproducibility (confidential elements withheld as Sponsor confidential program information); (iv) target-class generalizability favouring kinase and well-characterised enzyme targets; (v) no automated FTO check in the current workflow; (vi) docking-tool limitations for PPI interfaces; (vii) Scope of the “open-source” characterization: it applies to the computational discovery stack (library construction, docking, ADMET profiling, and scoring), all of which are free and open; the LLM is a substitutable orchestration component, not part of the compound-prioritization method. This study used a proprietary LLM API (Anthropic Claude) for that role, but the architecture is model-agnostic and can be operated with open-weight models (e.g., the Llama, Mistral, or Qwen families), so the blueprint entails no proprietary lock-in; a head-to-head evaluation of open-weight orchestration models is identified as future work.

#	Limitation	Description
1	No experimental validation	No experimental binding, selectivity, ADMET, cellular, or in vivo data were generated in this study. All compound-level claims are computational predictions requiring experimental confirmation.
2	No formal LLM efficiency comparison	LLM citation accuracy was measured by single-reviewer non-blinded verification of all available claim-citation pairs per case (n=8, 9, 10; per-pair verdicts in Supplementary S4); no controlled comparison against LLM-free workflow execution, no time-motion measurement, no inter-rater agreement.
3	Partial reproducibility	Case A target identity, target class, indication, structure PDB accession, library subset selection criteria, and all candidate compound SMILES are withheld as Sponsor confidential program information and not disclosed. Verbatim LLM prompt templates, prompt-chaining logic, and the agentic orchestration code are also withheld as confidential workflow components; their categorical structure is described in the Method section. Non-confidential elements (docking parameters, library construction rules, ADMET protocol) are fully specified in supplementary materials.

Table 4. Continued

#	Limitation	Description
4	Target-class generalizability	Empirical docking scoring functions (including the AutoDock Vina-derived scoring underlying GNINA) were trained predominantly on kinase, protease, and GPCR datasets; performance on other target classes (especially PPI interfaces) is more variable. Workflow is best-suited to kinase and well-characterised enzyme targets.
5	FTO and novelty	Literature-SAR scaffold restriction in Tier 1 enriches for documented chemotypes but selects against truly novel chemistry. Per-program FTO is performed routinely as a separate Sponsor workflow upstream of compound nomination and is not currently embedded in the agentic pipeline (see Freedom-to-Operate and Operational Practice, Discussion).
6	Custom-assay Tier 3 caveat	Tier 3 timeline and cost projections assume commercially available assay platforms (kinase, GPCR, ion channel, nuclear receptor). For less-established targets requiring custom assay development (e.g., PPI interfaces, undisclosed novel targets), Tier 3 timeline and cost can substantially exceed the projected figures.
7	PPI-interface docking limitation	The Case C structure is apo (no co-crystallised ligand) so a re-docking positive control was not possible for Case C; PPI interfaces are characteristically shallow and solvent-exposed, limiting rigid-receptor docking performance. Case C is framed as a highly exploratory stress test.

5.5. Interpretation of the Public Benchmark

The public benchmark reported above is a deliberately stringent test of docking-only prioritisation, and its outcome is informative. Accurate pose geometry (0.31 Å re-docking RMSD) did not translate into accurate affinity ranking. Several constraints bound the interpretation of these numbers. The benchmark was executed on CPU-only hardware using a single CNN model (crossdock_default2018) in rescoring mode rather than the five-model ensemble or CNN-guided refinement mode, both substantially more computationally demanding; enrichment statistics computed over 51 compounds, several of which depend on the top one to five ranked molecules, are correspondingly imprecise and should be read as indicative rather than definitive; and rigid-receptor docking models neither the flexible activation loop relevant to Case B nor the shallow, solvent-exposed interface of Case C. Accordingly, the result characterises the commodity scoring component under a constrained configuration, not the ceiling of the workflow. The practical implication is the design principle already embodied here: docking score functions as one weak filter within a multi-evidence funnel in which literature-derived target rationale, ADMET filtering, mechanism-based

interaction criteria and a mandatory human review gate carry the prioritisation decision. Strengthening the computational ranking itself - through ensemble CNN scoring or CNN-guided refinement on GPU hardware, MM-GBSA rescoring of top candidates (Genheden & Ryde, 2015), and target-specific machine-learning models trained on known actives where these exist - is the most direct route to improved screening performance and is identified here as the priority next step.

5.6. Implementation Note

Implementation note: the open-source workflow described here also forms the basis of a drug-discovery platform under development by Insivio. That platform is not yet commercially available; details of its planned deployment are intentionally omitted from this manuscript to keep the focus on the open-source scientific workflow. See the Conflict of Interest statement for the author's relationship with Insivio.

6. CONCLUSION

This paper describes a four-tier computational drug discovery pipeline built from open-source tools,

orchestrated by an LLM agent using a tiered Claude Opus 4 / Sonnet 4 selection, and demonstrated across three computational case studies spanning one confidential proprietary program (Case A), one well-characterised kinase target (Case B), and one PPI-interface stress test (Case C). Incurred LLM API cost across the three cases was \$1,450–\$1,800. Projected total discovery-phase cost including proposed Tier 3 work is \$3,450–\$10,800 (a lower bound; Tier 3 not executed in this study). LLM citation accuracy by PubMed verification was 88% (Case A, 7/8), 78% (Case B, 7/9), and 90% (Case C, 9/10); pooled 85% (23/27; descriptive count only, no pooled interval computed), with single non-blinded reviewer audit acknowledged as a limitation. The workflow is presented as a case-study-demonstrated open-source blueprint and is not a validated drug-discovery platform; compounds prioritised by this workflow are computational predictions requiring experimental confirmation in Tier 3 (proposed future work). The workflow is executable by a research group with computational-chemistry expertise at any institution with internet access and a standard multi-core workstation (no dedicated GPU required for the CPU-based docking described here).

Generative AI Use: Generative AI Use. AI writing assistance tools were used for grammar and style refinement. All scientific content, analysis, interpretation, and conclusions are the responsibility of the author.

Author Contributions: Budiwan Sumarsono: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review and editing, Visualization, Supervision, Project administration, Funding acquisition. The author confirms sole authorship and full responsibility for all content and declarations in this manuscript.

Funding: This research received no external funding. Computational resources, LLM API costs, and software licensing for this study were self-funded by the author through Insivio.

Conflict of Interest: The author is the Founder and CEO of Insivio, a company developing a drug-discovery platform based on the workflow described in this manuscript. The platform is currently under development and is not yet commercially available. The author has financial interests in Insivio. The study used open-source computational tools together with

a third-party commercial large-language-model API (Anthropic Claude); no proprietary Insivio software was required to generate the reported results, and the workflow can be reproduced from the published parameters without the Insivio platform. The in-house orchestration layer (LLM prompt templates and agentic code) is a substitutable component, withheld as confidential. The Insivio platform is discussed only in the Implementation Note within the Discussion section of the main manuscript.

Institutional Review Board Statement: Not applicable. This is a computational study with no human subjects, no patient data, and no animal experiments.

Informed Consent Statement: Not applicable. No human subjects were involved in this study.

Acknowledgments: The author thanks the reviewers and Associate Editor of the Journal of Artificial Intelligence and Knowledge Engineering for detailed and constructive feedback during the review process. The author also thanks Damar Yafi Sumarsono (IT Undergraduate Program, Monash University) for assistance with data cross-checking, manuscript editing, and figure preparation

Data and Code Availability: Data and Code Availability. Reproducibility of this workflow is partial. The non-confidential elements required to reproduce the computational pipeline — docking parameters, library construction rules, full ADMET filter cascade, scoring thresholds, and the ADMET protocol (Supplementary S2) — are provided in full. The categorical structure of the LLM-orchestration layer (role per tier, model selection, temperature, output schema, audit procedure) is described in the Method section. The verbatim LLM prompt templates, prompt-chaining logic, and the agentic orchestration code are withheld as Sponsor confidential workflow components and are not redistributed. Confidential program elements — the specific target identities and structure identifiers (including the Case A PDB accession), indications, the library subset selection criteria, and the SMILES of top-ranked candidate compounds in all three applied cases — are withheld as Sponsor confidential program information and are not disclosed in this manuscript. The three applied cases (A, B, C) are anonymized to target-class level (proprietary target; a well-characterised kinase; a PPI-interface target) pending provisional patent filing on the underlying programs; the identities will be released upon filing. To provide an independently

verifiable reproducibility check, a fully-specified public benchmark case (acetylcholinesterase, PDB 4EY7), with complete docking parameters and the executable script, is provided in Supplementary S3: the docking protocol recovers the crystallographic ligand pose to 0.31 Å RMSD (within the conventional 2.0 Å success gate), confirming that the docking protocol is correctly implemented and reproducible on public data. The non-confidential supplementary data and the public benchmark are provided in structured, machine-readable form consistent with FAIR data principles (Wilkinson et al., 2016).

Supplementary Materials

Supplementary S1 — Candidate Compound SMILES (redacted; structures withheld as Sponsor confidential program information. Library composition rules are provided in full).

Supplementary S2 — ADMET Protocol (SwissADME + ADMETlab 2.0 filter cascade with thresholds and pass criteria).

Supplementary S3 — Public benchmark case (acetylcholinesterase, PDB 4EY7): end-to-end reproducibility demonstration on a public target outside the Sponsor pipeline. The docking protocol recovers the crystallographic donepezil pose to 0.31 Å RMSD (≤ 2.0 Å success gate). Full parameters and the executable script are provided. Retrospective enrichment metrics (ROC-AUC, EF1%, EF10%, BEDROC) for docking score alone and with mechanism-based interaction filtering are reported.

Supplementary S4 — LLM citation-accuracy audit: the complete list of claim-citation pairs verified for each case, the accept/reject verdict and reason for each, the cited PMID and DOI, and the per-case and pooled Wilson confidence intervals. This is the underlying record for Table 2.

REFERENCES

- Boiko, D. A., MacKnight, R., Kline, B., & Gomes, G. (2023). Autonomous chemical research with large language models. *Nature*, 624(7992), 570–578. <https://doi.org/10.1038/s41586-023-06792-0>
- Bran, A. M., Cox, S., Schilter, O., Baldassari, C., White, A. D., & Schwaller, P. (2024). Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6, 525–535. <https://doi.org/10.1038/s42256-024-00832-8>
- Brown, L. D., Cai, T. T., & DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16(2), 101–133. <https://doi.org/10.1214/ss/1009213286>
- Comley, J. (2013, October). Outsourced kinase profiling services: Adding value to in-house kinase programmes. *Drug Discovery World*. <https://www.ddw-online.com/outsourced-kinase-profiling-services-adding-value-to-in-house-kinase-programmes-1631-201310/>
- Cook, D., Brown, D., Alexander, R., et al. (2014). Lessons learned from the fate of AstraZeneca's drug pipeline: A five-dimensional framework. *Nature Reviews Drug Discovery*, 13(6), 419–431. <https://doi.org/10.1038/nrd4309>
- Cross, J. B., Thompson, D. C., Rai, B. K., et al. (2009). Comparison of several molecular docking programs: Pose prediction and virtual screening accuracy. *Journal of Chemical Information and Modeling*, 49(6), 1455–1474. <https://doi.org/10.1021/ci900056c>
- Daina, A., Michielin, O., & Zoete, V. (2017). SwissADME: A free web tool to evaluate pharmacokinetics, drug-likeness and medicinal chemistry friendliness of small molecules. *Scientific Reports*, 7, 42717. <https://doi.org/10.1038/srep42717>
- Dara, S., Dhamercherla, S., Jadav, S. S., Babu, C. M., & Ahsan, M. J. (2022). Machine learning in drug discovery: A review. *Artificial Intelligence Review*, 55(3), 1947–1999. <https://doi.org/10.1007/s10462-021-10058-4>
- DiMasi, J. A., Grabowski, H. G., & Hansen, R. W. (2016). Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics*, 47, 20–33. <https://doi.org/10.1016/j.jhealeco.2016.01.012>
- Eberhardt, J., Santos-Martins, D., Tillack, A. F., & Forli, S. (2021). AutoDock Vina 1.2.0: New docking methods, expanded force field, and Python bindings. *Journal of Chemical Information and Modeling*, 61(8), 3891–3898. <https://doi.org/10.1021/acs.jcim.1c00203>
- Gaulton, A., Hersey, A., Nowotka, M., et al. (2017). The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1), D945–D954. <https://doi.org/10.1093/nar/gkw1074>
- Genheden, S., & Ryde, U. (2015). The MM/PBSA and MM/GBSA methods to estimate ligand-

- binding affinities. *Expert Opinion on Drug Discovery*, 10(5), 449–461. <https://doi.org/10.1517/17460441.2015.1032936>
- Irwin, J. J., Shoichet, B. K., Mysinger, M. M., et al. (2009). Automated docking screens: A feasibility study. *Journal of Medicinal Chemistry*, 52(18), 5712–5726. <https://doi.org/10.1021/jm9006966>
- Jayatunga, M. K. P., Xie, W., Ruder, L., Schulze, U., & Meier, C. (2022). AI in small-molecule drug discovery: A coming wave? *Nature Reviews Drug Discovery*, 21(3), 175–176. <https://doi.org/10.1038/d41573-022-00025-1>
- Jorgensen, W. L. (2004). The many roles of computation in drug discovery. *Science*, 303(5665), 1813–1818. <https://doi.org/10.1126/science.1096361>
- Karaman, M. W., Herrgard, S., Treiber, D. K., et al. (2008). A quantitative analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 26(1), 127–132. <https://doi.org/10.1038/nbt1358>
- McNutt, A. T., Francoeur, P., Aggarwal, R., et al. (2021). GNINA 1.0: Molecular docking with deep learning. *Journal of Cheminformatics*, 13(1), 43. <https://doi.org/10.1186/s13321-021-00522-2>
- Mysinger, M. M., Carchia, M., Irwin, J. J., & Shoichet, B. K. (2012). Directory of Useful Decoys, Enhanced (DUD-E): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14), 6582–6594. <https://doi.org/10.1021/jm300687e>
- Schrödinger, Inc. (2023). Annual report (Form 10-K) for the fiscal year ended December 31, 2023. United States Securities and Exchange Commission.
- Sewak, M. (2026). The Ascent of Machine Cognition: A Systematic Review of Large Reasoning Models. *Journal of Artificial Intelligence and Knowledge Engineering*, 1(1), 1. <https://doi.org/10.37819/Jaike.25.0010>
- Truchon, J.-F., & Bayly, C. I. (2007). Evaluating virtual screening methods: Good and bad metrics for the “early recognition” problem. *Journal of Chemical Information and Modeling*, 47(2), 488–508. <https://doi.org/10.1021/ci600426e>
- Wang, Z., Sun, H., Yao, X., et al. (2016). Comprehensive evaluation of ten docking programs on a diverse set of protein–ligand complexes: The prediction accuracy of sampling power and scoring power. *Physical Chemistry Chemical Physics*, 18(18), 12964–12975. <https://doi.org/10.1039/C6CP01555G>
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>
- Xiong, G., Wu, Z., Yi, J., et al. (2021). ADMETlab 2.0: An integrated online platform for accurate and comprehensive predictions of ADMET properties. *Nucleic Acids Research*, 49(W1), W5–W14. <https://doi.org/10.1093/nar/gkab255>



Publisher's note: Eurasia Academic Publishing Group (EAPG) and AAKS remain neutral about jurisdictional claims in published maps and institutional affiliations.

Open Access. This article is licensed under a Creative Commons Attribution-NoDerivatives 4.0 International (CC BY-ND 4.0) licence, which permits copy and redistribute the material in any medium or format for any purpose, even commercially. The licensor cannot revoke these freedoms as long as you follow the licence terms. Under the following terms you must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorsed you or your use. If you remix, transform, or build upon the material, you may not distribute the modified material. To view a copy of this license, visit <https://creativecommons.org/licenses/by-nd/4.0/>.

SUPPLEMENTARY MATERIALS

Supplementary S1. Candidate Compound SMILES (redacted; structures withheld as Sponsor confidential program information. Library composition rules are provided in full)

NOTICE — Confidential Sponsor program information. Per the Sponsor's confidentiality policy and reviewer-flagged disclosure scope, full SMILES strings, canonical InChIKeys, and synthesis-enabling compound identifiers for the top-ranked hits in all three cases are withheld as Sponsor confidential

program information and will not be disclosed in this manuscript or supplement. For Case A, the target identity, target class, indication, structure PDB accession, and the library subset selection criteria are likewise withheld. This supplement provides the public-database anchors and audit metadata needed to reproduce the workflow design (where the target is public) without enabling re-derivation of the proprietary hits.

1. Per-case summary (non-enabling)

Case	Target	Scaffold class	Public anchor
A	Proprietary therapeutic target (target class and indication withheld)	Withheld as Sponsor confidential program information	Not disclosed (experimental co-crystal structure; PDB accession withheld)
B	Well-characterised kinase (identity anonymized)	Kinase chemotype (identity anonymized pending provisional filing)	Public co-crystal structure (identifier anonymized)
C	PPI-interface target (identity anonymized)	Small-molecule PPI-interface binder (apo groove)	Public apo structure (identifier anonymized)

2. Workflow audit anchors (reproducible)

For each case, the workflow retains: (i) the locked Tier-1 LLM prompt and tool registry hash (held internally as a confidential workflow component; categorical structure described in the Method section of the main manuscript), (ii) the public structure (PDB) and grid-box definition used for docking, (iii) the GNINA v1.1 CNN-rescored docking-score distributions (reported as rank-order scores, not ΔG), and (iv) the ADMET surrogate panel run (Supplementary S2). These artifacts are sufficient for an independent team to reproduce the screening protocol on their own compound library.

3. Reproducibility scope

Reproducibility is partial by design. Public components — docking protocol, grid boxes, ADMET model versions, GNINA CNN-rescoring parameters — are released (Supplementary S2). The verbatim LLM prompt templates, prompt-chaining logic, and the agentic orchestration code are withheld as Sponsor confidential workflow components; their categorical structure is described in the Method section of the main manuscript. Confidential program components — top-hit SMILES, per-pose docking logs, custom compound library, and any unpublished SAR — are withheld as Sponsor

confidential program information. This trade-off is disclosed explicitly in the Limitations subsection of the Discussion and is consistent with the disclosure scope discussed in the R2 response letter.

4. Verification access for editors and reviewers

End-to-end reproducibility is provided by a fully-specified public benchmark case (Supplementary S3) on a target outside the sponsor's pipeline. The three applied cases (A, B, C) are anonymized to target-class level; their specific target identities and structure identifiers, together with candidate-compound SMILES and the Case A model coordinates, will be made available to the Editor upon filing of the corresponding provisional patent applications (anticipated in 2026). Enabling materials are not released during peer review, as they are the subject of pending intellectual-property protection.

Supplementary S2. ADMET Protocol (SwissADME + ADMETlab 2.0 filter cascade with thresholds and pass criteria)

Docking Protocol Parameters and ADMET Filtering Thresholds

All tools are open-source unless noted. Versions as of April 2026. The three applied cases (A, B, C) are anonymized to target-class level pending provisional patent filing; case-specific structure identifiers and any Sponsor confidential program information are withheld. This supplement specifies the executed computational protocol (Tier 1 library construction and ADMET filtering; Tier 2 GNINA docking with CNN rescoring). MM-GBSA rescoring and free-energy-perturbation calculations are not part of the reported workflow; the reported computational

binding estimates are GNINA CNN-rescored docking scores used for rank-order triage only.

1. Molecular Docking Protocol (Tier 2)

Docking used a single engine (GNINA v1.1): an AutoDock Vina-derived empirical search followed by CNN affinity rescoring of final poses (--cnn_scoring rescore). Candidates are ranked by GNINA CNN-rescored affinity (a unitless, higher-is-better predicted-affinity score), used for ranking only; the empirical (Vina-derived) term is reported in kcal/mol but is not the ranking quantity and is not interpreted as a free energy of binding (not ΔG).

Parameter	Value / Description
Docking software	GNINA v1.1 (AutoDock Vina-derived empirical search + CNN affinity rescoring, --cnn_scoring rescore)
Advancement rule	Compounds advanced if ranked in the top 30% by GNINA CNN-rescored affinity
Binding box	22.5 x 22.5 x 22.5 Å, centred per case (see box definition below)
Exhaustiveness	32 (Case A); 16 (Cases B, C). AutoDock Vina default is 8.
Binding modes	9 per ligand
Independent runs	Fixed random seed 42; 9 binding modes retained per ligand
Reported value	GNINA CNN-rescored affinity of the top-ranked pose (single run; fixed seed 42)
Protein preparation	Hydrogens added (Open Babel 3.1.1); Gasteiger charges; waters removed
Ligand preparation	RDKit ETKDG 3D conformer generation; MMFF94 optimisation; 50 conformers

1a. Per-case docking-box definition and re-docking control

Case	Docking-box definition	Re-docking positive control
A (proprietary; public experimental co-crystal structure)	Box centred on the co-crystallised ligand resolved in the experimental structure, which defines the primary ligand-binding pocket.	The native co-crystallised ligand was re-docked into its own experimental structure as a positive-control check before novel-compound docking; it was recovered in the crystallographic binding site with reproducible top-ranked docking scores across independent runs (a quantitative pose-recovery RMSD benchmark is provided for the public case in S3). The specific PDB accession is withheld as Sponsor confidential program information.
B (kinase; public co-crystal)	Box centred on the ATP-site inhibitor-binding region defined by the co-crystallised inhibitor in the public structure.	The co-crystallised inhibitor was re-docked solely as a box-placement check (binding-site localisation confirmed); it is not a rigorous re-docking positive control.
C (PPI interface; public apo)	Box centred on the partner-binding (PPI) groove using residue contacts identified directly on the apo structure (homolog-free, apo-derived box).	Not possible: the structure is apo (no co-crystallised ligand). Absence of a re-docking positive control contributes to the highly exploratory framing of Case C.

2. Tier 1 ADMET Filtering Thresholds

Applied sequentially; a compound failing any hard criterion is eliminated before docking.

Physicochemical and drug-likeness endpoints computed with RDKit and SwissADME; predictive toxicity and pharmacokinetic endpoints with ADMETlab 2.0.

Filter	Threshold / Criterion	Tool
Lipinski Rule of Five	MW ≤ 500, LogP ≤ 5, HBD ≤ 5, HBA ≤ 10; max 1 violation	RDKit / SwissADME
Veber oral bioavailability	TPSA ≤ 140 Å ² , rotatable bonds ≤ 10	RDKit / SwissADME
PAINS exclusion	Removal of pan-assay interference scaffolds (PAINS-A, B, C)	RDKit FilterCatalog
Brenk unwanted fragments	Exclusion of reactive/toxic substructure patterns	RDKit FilterCatalog
hERG cardiotoxicity	Predicted probability < 0.5	ADMETlab 2.0
Ames mutagenicity	Predicted probability < 0.5	ADMETlab 2.0

Filter	Continued	
	Threshold / Criterion	Tool
CYP3A4 inhibition	Predicted probability < 0.5	ADMETlab 2.0
CYP2D6 inhibition	Predicted probability < 0.5	ADMETlab 2.0
Aqueous solubility	LogS > -5 preferred; flagged if < -6	ADMETlab 2.0
BBB penetration	Computed for all compounds; applied as a hard filter on a per-program basis	ADMETlab 2.0
ADMET composite score	>= 60 / 100 required to advance	Sponsor internal scoring

3. Compound Library Construction (Tier 1)

Parameter	Value / Description
Source	ZINC20 (Irwin et al., J. Chem. Inf. Model. 2020)
Case A	Target-class-appropriate ZINC20 subset; specific subset and selection criteria withheld as Sponsor confidential program information
Case B	ZINC20 kinase-focused subset (>=40% kinase domain sequence identity to the target)
Case C	ZINC20 general lead-like subset
Tier 1 filters	Lipinski Rule-of-Five, PAINS exclusion, reactive-group (Brenk) exclusion, TPSA cap (see Section 2)

3a. Library reduction by case

Case A is reported at target-class level; its initial library size is disclosed but subset criteria are withheld.

Case	Initial library	Post Tier-1 ADMET	Docked	Advanced to proposed Tier 3
A	~400,000	1,847	200	10
B	388,000	2,103	175	10
C	154,000	-	154	10

Hardware: all docking was executed on CPU-only consumer hardware; no GPU was used. A single docking engine (GNINA v1.1) was used for all reported results; AutoDock Vina and Smina were not used to generate any result reported in this study.

Case C: the general lead-like library was carried into docking (154 candidates docked); the 10 top-ranked compounds are framed as exploratory only (see Result: Case C).

Supplementary S3. Public benchmark case (acetylcholinesterase, PDB 4EY7): end-to-end reproducibility demonstration on a public target

outside the Sponsor pipeline. The docking protocol recovers the crystallographic donepezil pose to 0.31 Å RMSD (≤ 2.0 Å success gate). Full parameters and the executable script are provided. Retrospective enrichment metrics (ROC-AUC, EF1%, EF10%, BEDROC) for docking score alone and with mechanism-based interaction filtering are reported

Public Benchmark Case: Reproducibility Demonstration

S3.1 Purpose

This benchmark demonstrates that the docking-and-scoring component of the workflow described in the manuscript can be reproduced end-to-end, independently, on a fully public target that is not part of the sponsor's pipeline. Because the three applied case studies have anonymized targets (pending provisional patent filing) and no experimental ground truth, this public benchmark provides an objective, self-contained reproducibility check that any reader can execute. The benchmark evaluates reproducibility of the docking protocol (its ability to recover a known crystallographic binding pose); it is not a virtual-screening accuracy claim. Consistent with the manuscript, docking scores are treated as ranking outputs, not affinities, and all prioritised compounds require experimental confirmation in Tier 3.

S3.2 Target and data (all public)

- Target: human acetylcholinesterase (AChE).
- Structure: PDB 4EY7 (AChE in complex with the approved inhibitor donepezil; ligand HET

code E20). Chain A and a single co-crystallised donepezil copy were retained.

- All inputs (PDB coordinates, ligand) are public; this supplement contains no sponsor confidential program information.

S3.3 Method (identical Tier-2 docking configuration; Tier-1 ADMET simplified to a public-tool subset)

Receptor and ligand were prepared with Open Babel (protonation at pH 7.4; Gasteiger charges). Docking used GNINA v1.1 with CNN rescoring (--cnn_scoring rescore, single CNN model crossdock_default2018, CPU), exhaustiveness 16, a 22.5 x 22.5 x 22.5 Å box centred on the co-crystallised ligand centroid, and fixed random seed 42. GNINA is a distinct program from AutoDock Vina and Smina: its search and scoring are not interchangeable with those tools, and no result reported here was produced with AutoDock Vina or Smina. For the positive control, the co-crystallised ligand (donepezil) was re-docked and the top-pose heavy-atom RMSD to the crystallographic pose was computed with obrms (0.31 Å; success gate ≤ 2.0 Å).

S3.4 Result – re-docking positive control

Metric	Result
Re-docking positive control (top-pose heavy-atom RMSD vs crystal)	0.31 Å
Acceptance gate	≤ 2.0 Å — PASS
Docking engine / scoring	GNINA v1.1, CNN rescoring (--cnn_scoring rescore, crossdock_default2018; CPU)
Box / search	22.5 Å cubic box centred on crystal-ligand centroid; exhaustiveness 16; seed 42

The top re-docked pose reproduces the crystallographic donepezil binding mode to 0.31 Å RMSD, well within the conventional 2.0 Å success threshold. This confirms that the docking protocol, exactly as specified, recovers a known binding pose on a public target — an objective, independently verifiable demonstration that the workflow is correctly implemented and reproducible.

Retrospective enrichment. Actives were high-confidence AChE inhibitors from ChEMBL (ChEMBL220, pChEMBL ≥ 7); decoys were property-matched DUD-E decoys (aces). Across 51 scored compounds (15 actives, 36 decoys), ranking by docking score alone gave ROC-AUC 0.23, EF1%

0.0x, EF10% 0.0x and BEDROC($\alpha = 20$) 0.02. Adding a mechanism-based interaction criterion (top pose contacting the conserved catalytic-gorge anchor tryptophans Trp86/Trp286, matching the shipped filter script; applied identically to actives and decoys) gave ROC-AUC 0.44, EF1% 3.4x, EF10% 1.4x, BEDROC 0.44 and top-10 hit-rate 0.20. Because this criterion encodes prior knowledge of the AChE binding mode, its enrichment gain is illustrative rather than target-agnostic, and EF1% rests on a single top-ranked compound. Limitations: single CNN model in rescoring mode on CPU (not the five-model ensemble or CNN-guided refinement); $n = 51$, so metrics resting on the top 1-5 ranked

molecules are imprecise; GNINA CNN affinity favours larger, more lipophilic ligands and DUD-E decoys carry documented property-matching biases. These figures characterise the commodity scoring component under a constrained configuration, not the workflow as a whole.

S3.5 Reproduce it

Parameter-locked script is provided: S3_benchmark_r3par_wsl.py (local GNINA/CPU), together with the interaction-filter script and the per-compound results CSV. No GPU and no paid compute are required.

Local (WSL/Linux, after installing GNINA + Open Babel):

```
python3 S3_benchmark_wsl.py # downloads 4EY7, prepares, re-docks, reports RMSD
```

Colab (free, no local GPU):

```
!apt-get -qq install -y openbabel; pip -q install rdkit requests
```

All data used are public (PDB, ChEMBL); this benchmark contains no sponsor confidential program information.

within this supplement. Both scripts run on CPU with GNINA v1.1 and Open Babel; no GPU and no paid compute are required. The per-compound results file (S3_r3_results.csv) accompanies this supplement.

S3.6 Reproducibility scripts (full source)

The complete, parameter-locked source for the public AChE/4EY7 benchmark is reproduced below so that it is available to reviewers and readers directly

S3.6.1 Benchmark script – S3_benchmark_r3par_wsl.py

```
#!/usr/bin/env python3
"""R3 AChE/4EY7 enrichment benchmark – GNINA cnn_scoring=rescore, PARALLEL (CPU).
Matches corrected method (GNINA v1.1, --cnn_scoring=rescore). Docks ligands
concurrently (default 6 workers x 2 threads). Prints PROGRESS lines "PROGRESS d/N".
Reports re-dock RMSD, ROC-AUC, EF1%, EF10%, BEDROC(a=20), hit-rate top-N.
Run: python3 S3_benchmark_r3par_wsl.py --n_actives 15 --n_decoys 30 --exh 16 --workers 6
"""
import os, sys, subprocess, shutil, urllib.request, csv, math, argparse, time
from concurrent.futures import ThreadPoolExecutor, as_completed
from rdkit import Chem
from rdkit.Chem import AllChem, Descriptors, Crippen, rdMolDescriptors, FilterCatalog
AP = argparse.ArgumentParser()
AP.add_argument("--n_actives", type=int, default=20)
AP.add_argument("--n_decoys", type=int, default=40)
AP.add_argument("--exh", type=int, default=16)
AP.add_argument("--workers", type=int, default=6)
AP.add_argument("--cpu", type=int, default=2)
AP.add_argument("--box", type=float, default=22.5) # fixed cubic box (matches case-study S2
protocol)
AP.add_argument("--outdir", default=os.path.expanduser("~/S3_r3b_out"))
A = AP.parse_args()
os.makedirs(A.outdir, exist_ok=True)
GNINA = shutil.which("gnina"); OBABEL = shutil.which("obabel")
assert GNINA and OBABEL, "gnina/obabel not on PATH"
# single CNN model (not the 5-model ensemble) => ~5x less RAM, avoids OOM on 8GB
EXTRA = ["--cnn_scoring", "rescore", "--cnn", "crossdock_default2018"]
def run(cmd): return subprocess.run(cmd, capture_output=True, text=True)
# 1. structure
pdb = os.path.join(A.outdir, "4EY7.pdb")
if not os.path.exists(pdb):
    urllib.request.urlretrieve("https://files.rcsb.org/download/4EY7.pdb", pdb)
```

```

rec_pdb = os.path.join(A.outdir, "receptor.pdb"); lig_pdb = os.path.join(A.outdir, "xtal_
lig.pdb")
with open(pdb) as f, open(rec_pdb, "w") as fr, open(lig_pdb, "w") as fl:
for ln in f:
if ln.startswith("ATOM") and ln[21] == "A": fr.write(ln)
elif ln.startswith("HETATM") and ln[17:20].strip() == "E20" and ln[21] == "A": fl.write(ln)
rec_pdbqt = os.path.join(A.outdir, "receptor.pdbqt"); lig_pdbqt = os.path.join(A.outdir,
"xtal_lig.pdbqt")
run([OBABEL, rec_pdb, "-xr", "-O", rec_pdbqt, "-p", "7.4"])
run([OBABEL, lig_pdb, "-O", lig_pdbqt, "-p", "7.4", "--partialcharge", "gasteiger"])
# fixed cubic box centred on crystal-ligand centroid (matches case-study 22.5 Å box)
xs=[];ys=[];zs=[]
for ln in open(lig_pdb):
if ln.startswith("HETATM"):
xs.append(float(ln[30:38])); ys.append(float(ln[38:46])); zs.append(float(ln[46:54]))
cx,cy,cz=sum(xs)/len(xs),sum(ys)/len(ys),sum(zs)/len(zs)
BOX=["--center_x",f"{cx:.3f}", "--center_y",f"{cy:.3f}", "--center_z",f"{cz:.3f}",
"--size_x",str(A.box), "--size_y",str(A.box), "--size_z",str(A.box)]
print(f"[1] receptor+ligand prepared; box {A.box} Å centred ({cx:.1f},{cy:.1f},{cz:.1f})",
flush=True)
# 2. re-dock control
redock = os.path.join(A.outdir, "redock.sdf")
run([GNINA, "-r", rec_pdbqt, "-l", lig_pdbqt, *BOX,
"--exhaustiveness", str(A.exh), "--num_modes", "5", "-o", redock, "--seed", "42", "--cpu",
str(A.cpu)] + EXTRA)
xtal_sdf = os.path.join(A.outdir, "xtal_lig.sdf"); run([OBABEL, lig_pdb, "-O", xtal_sdf,
"-p", "7.4"])
rms = run(["obrms", xtal_sdf, redock])
try: rmsd = float(rms.stdout.strip().split()[-1])
except: rmsd = rms.stdout.strip() or rms.stderr.strip()
print(f"[2] Re-docking RMSD: {rmsd} Å (gate <=2.0)", flush=True)
# 3. actives + decoys
import requests
def chembl_actives(n):
# curated: high-confidence potent AChE inhibitors (pChEMBL >= 7, i.e. IC50/Ki <= 100 nM),
# de-duplicated by canonical SMILES, ordered by potency (most potent first)
cand=[]
for offset in (0,1000):
try:
js = requests.get("https://www.ebi.ac.uk/chembl/api/data/activity.json?"
"target_chembl_id=CHEMBL220&pchembl_value__gte=7&standard_type__in=IC50,Ki"
f"&limit=1000&offset={offset}", timeout=90).json()
except Exception as e:
print(" ChEMBL page failed:", e); break
acts=js.get("activities", [])
for a in acts:
smi=a.get("canonical_smiles"); pv=a.get("pchembl_value")
if smi and pv and "." not in smi:
try: cand.append((float(pv), smi))
except: pass
if len(acts)<1000: break
cand.sort(key=lambda x:-x[0])
out, seen = [], set()
for pv,smi in cand:
if smi not in seen: seen.add(smi); out.append(smi)
print(f" actives: curated pChEMBL>=7, {len(out)} unique available", flush=True)
return out[:n]
def get_decoys(n):
import random

```

```

try:
txt = requests.get("https://dude.docking.org/targets/aces/decoys_final.ism", timeout=60).
text
smis = list(dict.fromkeys(l.split()[0] for l in txt.splitlines() if l.strip()))
if len(smis) >= n:
random.seed(42); random.shuffle(smis); print(" decoys: DUD-E aces", flush=True); return
smis[:n]
except Exception as e:
print(" DUD-E failed:", e, flush=True)
js = requests.get("https://www.ebi.ac.uk/chembl/api/data/molecule.json?"
"molecule_properties__mw_freebase__lte=500&molecule_properties__mw_freebase__
gte=200&limit=1000", timeout=90).json()
out = list(dict.fromkeys((m.get("molecule_structures") or {}).get("canonical_smiles") for m
in js.get("molecules", [])))
out = [s for s in out if s and "." not in s]
print(" decoys: ChEMBL random drug-like (DUD-E unavailable)", flush=True); return out[:n]
actives = chembl_actives(A.n_actives); aset = set(actives)
decoys = [s for s in get_decoys(A.n_decoys + 30) if s not in aset][:A.n_decoys]
print(f"[3] actives={len(actives)} decoys={len(decoys)}", flush=True)
pains = FilterCatalog.FilterCatalogParams(); pains.AddCatalog(FilterCatalog.
FilterCatalogParams.FilterCatalogs.PAINS)
CAT = FilterCatalog.FilterCatalog(pains)
def tier1(m):
mw, logp=Descriptors.MolWt(m),Crippen.MolLogP(m)
hbd,hba,tpsa=rdMolDescriptors.CalcNumHBD(m),rdMolDescriptors.
CalcNumHBA(m),rdMolDescriptors.CalcTPSA(m)
return (((mw<=500)+(logp<=5)+(hbd<=5)+(hba<=10)>=3) and tpsa<=140 and not CAT.HasMatch(m)),
dict(MW=round(mw,1),logP=round(logp,2),TPSA=round(tpsa,1))
# prep all ligand SDFs (fast, serial)
jobs=[]
allmols=[("active",i,s) for i,s in enumerate(actives)]+[("decoy",i,s) for i,s in
enumerate(decoys)]
rows_bad=[]
for tag,i,smi in allmols:
m=Chem.MolFromSmiles(smi)
if m is None: rows_bad.
append(dict(tag=tag,tier1=False,cnn=None,emp=None,smiles=smi,MW=None,logP=None,TPSA=None));
continue
ok,props=tier1(m)
if not ok: rows_bad.append(dict(tag=tag,tier1=False,cnn=None,emp=None,smiles=smi,**props));
continue
mh=Chem.AddHs(m)
if AllChem.EmbedMolecule(mh,randomSeed=0xf00d)!=0: rows_bad.
append(dict(tag=tag,tier1=True,cnn=None,emp=None,smiles=smi,**props)); continue
AllChem.MMFFOptimizeMolecule(mh)
sdf=os.path.join(A.outdir,f"l_{tag}_{i}.sdf"); Chem.MolToMolFile(mh,sdf)
jobs.append((tag,i,smi,props,sdf))
N_JOBS=len(jobs); print(f"[4] docking {N_JOBS} ligands, {A.workers} workers x {A.cpu} cpu,
exh {A.exh}", flush=True)
def parse_scores(o):
cnn=emp=None
try:
if os.path.exists(o) and os.path.getsize(o)>0:
for mol in Chem.SDMolSupplier(o):
if mol is None: continue
if mol.HasProp("CNNAffinity"): cnn=float(mol.GetProp("CNNAffinity"))
if mol.HasProp("minimizedAffinity"): emp=float(mol.GetProp("minimizedAffinity"))
break
except Exception: pass
return cnn,emp
def dock(job):

```

```

tag,i,smi,props,sdf=job
o=sdf.replace(".sdf","_d.sdf")
try:
cnn,emp=parse_scores(o) # RESUME: reuse a valid prior result
if cnn is None:
run([GNINA,"-r",rec_pdbqt,"-l",sdf,*BOX,
"--exhaustiveness",str(A.exh),"--num_modes","3","-o",o,"--seed","42","--cpu",str(A.
cpu)]+EXTRA)
cnn,emp=parse_scores(o)
except Exception as e:
print(f" WARN dock {tag}_{i}: {e}", flush=True); cnn=emp=None
return dict(tag=tag,tier1=True,cnn=cnn,emp=emp,smiles=smi,**props)
t0=time.time(); rows=list(rows_bad); done=0
with ThreadPoolExecutor(max_workers=A.workers) as ex:
futs={ex.submit(dock,j):j for j in jobs}
for fut in as_completed(futs):
rows.append(fut.result()); done+=1
el=time.time()-t0; eta=el/done*(N_JOBS-done)
print(f"PROGRESS {done}/{N_JOBS} ({100*done/N_JOBS}%) elapsed={el/60:.1f}m
eta={eta/60:.1f}m", flush=True)
# 5. enrichment (rank by CNN affinity, higher=better)
sc=[r for r in rows if r and r["cnn"] is not None]; sc.sort(key=lambda r:-r["cnn"])
N=len(sc); nA=sum(r["tag"]=="active" for r in sc)
labels=[1 if r["tag"]=="active" else 0 for r in sc]
def ef(f):
k=max(1,int(N*f)); return (sum(labels[:k])/k)/(nA/N) if nA else 0
def hit(k): k=min(k,N); return sum(labels[:k])/k if k else 0
def auc():
As=[r["cnn"] for r in sc if r["tag"]=="active"]; Ds=[r["cnn"] for r in sc if
r["tag"]=="decoy"]
return sum((a>d)+0.5*(a==d) for a in As for d in Ds)/(len(As)*len(Ds)) if As and Ds else
float("nan")
def bedroc(al=20.0):
if nA in (0,N): return float("nan")
ra=nA/N; s=sum(math.exp(-al*(i+1)/N) for i,l in enumerate(labels) if l)
rie=(s/nA)/((1.0/N)*(1-math.exp(-al))/(math.exp(al/N)-1))
return rie*(ra*math.sinh(al/2)/(math.cosh(al/2)-math.cosh(al/2-al*ra)))+1.0/(1-math.
exp(al*(1-ra)))
with open(os.path.join(A.outdir,"S3_r3_results.csv"),"w",newline="") as f:
w=csv.DictWriter(f,fieldnames=["tag","tier1","cnn","emp","MW","logP","TPSA","smiles"]);
w.writeheader()
[w.writerow({k:r.get(k) for k in w.fieldnames}) for r in rows if r]
print("\n==== R3 BENCHMARK (AChE/4EY7, GNINA cnn_scoring=rescore, exh %d) ===== % A.exh)
print(f"Re-docking RMSD : {rmsd} A")
print(f"Scored : {N} ({nA} actives, {N-nA} decoys)")
print(f"ROC-AUC : {auc():.3f} EF1% : {ef(0.01):.1f}x EF10% : {ef(0.10):.1f}x BEDROC(a=20) :
{bedroc():.3f}")
print(f"Hit-rate top-5 : {hit(5):.2f} top-10 : {hit(10):.2f}")
print(f"Total docking wall time: {(time.time()-t0)/60:.1f} min")
print("DONE_MARKER")

```

S3.6.2 Interaction-filter script – interaction_filter.py

```

#!/usr/bin/env python3
"""Post-docking interaction filter for the AChE/4EY7 benchmark.
Re-uses existing docked poses in --outdir (no re-docking). Keeps molecules whose
best pose contacts the catalytic-gorge aromatic anchors (Trp86 / Trp286), scores
how many key gorge residues each contacts, then re-ranks and recomputes enrichment.
Mechanism-based, applied identically to actives and decoys.

```

```
"""
```

```
import os, csv, math, glob, argparse
from rdkit import Chem
AP=argparse.ArgumentParser()
AP.add_argument("--outdir", default=os.path.expanduser("~/S3_r3b_out"))
AP.add_argument("--cutoff", type=float, default=4.0) # heavy-atom contact distance (A)
```

```
A=AP.parse_args()
# human AChE (4EY7) catalytic-gorge residues (resName,resSeq)
ANCHORS={"TRP","86"},("TRP","286")} # gate: pi-stacking anchors (CAS / PAS)
KEY={"TRP","86"},("TRP","286"),("TYR","337"),("PHE","338"),("TYR","124"),
("TYR","72"),("HIS","447"),("SER","203"),("GLU","334"),("GLU","202"),("TYR","341")}
# parse receptor heavy atoms for key residues
```

```
rec_pdb=os.path.join(A.outdir,"receptor.pdb")
res_atoms={} # (resName,resSeq) -> list[(x,y,z)]
for ln in open(rec_pdb):
    if not ln.startswith("ATOM"): continue
    rn=ln[17:20].strip(); rs=ln[22:26].strip(); el=ln[76:78].strip() or ln[12:16].strip()[0]
    if el=="H": continue
```

```
key=(rn,rs)
if key in KEY:
    res_atoms.setdefault(key,[]).append((float(ln[30:38]),float(ln[38:46]),float(ln[46:54])))
found=sorted(f"{k[0]}{k[1]}{len(v)}" for k,v in res_atoms.items())
print("key residues found:", " ", ".join(found) or "NONE")
missing=[f"{r}{s}" for (r,s) in KEY if (r,s) not in res_atoms]
```

```
if missing: print("MISSING:", " ", ".join(missing))
def pose_coords(sdf):
    if not (os.path.exists(sdf) and os.path.getsize(sdf)>0): return None,None,None
    for mol in Chem.SDMolSupplier(sdf, removeHs=True):
        if mol is None: continue
        conf=mol.GetConformer()
```

```
xyz=[(conf.GetAtomPosition(i).x,conf.GetAtomPosition(i).y,conf.GetAtomPosition(i).z) for i
in range(mol.GetNumAtoms())]
cnn=float(mol.GetProp("CNNAffinity")) if mol.HasProp("CNNAffinity") else None
emp=float(mol.GetProp("minimizedAffinity")) if mol.HasProp("minimizedAffinity") else None
return xyz,cnn,emp
return None,None,None
```

```
def contacts(xyz):
    hit=set()
    for key,atoms in res_atoms.items():
        for (lx,ly,lz) in xyz:
            done=False
            for (rx,ry,rz) in atoms:
                if (lx-rx)**2+(ly-ry)**2+(lz-rz)**2 <= A.cutoff**2:
```

```
                hit.add(key); done=True; break
            if done: break
    return hit
# read ALL docked poses via glob (indices have gaps where ligands failed prep)
rows=[]
for sdf in sorted(glob.glob(os.path.join(A.outdir,"l_*_d.sdf"))):
```

```
    tag = "active" if os.path.basename(sdf).startswith("l_active_") else "decoy"
    xyz,cnn,emp=pose_coords(sdf)
    if xyz is not None:
        hit=contacts(xyz)
        rows.append(dict(tag=tag, cnn=cnn, emp=emp, nkey=len(hit), gate=bool(hit & ANCHORS)))
    nA=sum(r["tag"]=="active" for r in rows); N=len(rows)
```

```
gA=sum(1 for r in rows if r["tag"]=="active" and r["gate"])
gD=sum(1 for r in rows if r["tag"]=="decoy" and r["gate"])
print(f"\nligands with poses: {N} ({nA} actives, {N-nA} decoys)")
print(f"gate-pass (contacts Trp86/Trp286): actives {gA}/{nA}, decoys {gD}/{N-nA}")
def metrics(name, keyfn, higher=True):
    sc=[r for r in rows if keyfn(r) is not None]
```

```
sc.sort(key=lambda r:(-keyfn(r) if higher else keyfn(r)))
n=len(sc); na=sum(r["tag"]=="active" for r in sc)
lab=[1 if r["tag"]=="active" else 0 for r in sc]
def ef(f):
k=max(1,int(n*f)); return (sum(lab[:k])/k)/(na/n) if na else 0
def hit(k): k=min(k,n); return sum(lab[:k])/k if k else 0

A_=[keyfn(r) for r in sc if r["tag"]=="active"]; D_=[keyfn(r) for r in sc if
r["tag"]=="decoy"]
if higher: auc=sum((a>d)+0.5*(a==d) for a in A_ for d in D_)/(len(A_)*len(D_)) if A_ and D_
else float("nan")
else: auc=sum((a<d)+0.5*(a==d) for a in A_ for d in D_)/(len(A_)*len(D_)) if A_ and D_ else
float("nan")
if na in (0,n): bed=float("nan")
else:
al=20.0; ra=na/n; s=sum(math.exp(-al*(i+1)/n) for i,l in enumerate(lab) if l)
rie=(s/na)/((1.0/n)*(1-math.exp(-al))/(math.exp(al/n)-1))
bed=rie*(ra*math.sinh(al/2)/(math.cosh(al/2)-math.cosh(al/2-al*ra)))+1.0/(1-math.exp(al*(1-
ra)))
print(f" {name:28s} ROC-AUC={auc:.3f} EF1%={ef(0.01):.1f}x EF10%={ef(0.10):.1f}x
BEDROC={bed:.3f} top10={hit(10):.2f}")
# combined score: gate-pass first, then more key contacts, then CNN affinity
def combo(r):
return (1000 if r["gate"] else 0) + 10*r["nkey"] + (r["cnn"] or 0)
print("\n===== ENRICHMENT =====")
print("BEFORE filter:")
metrics("CNN affinity only", lambda r:r["cnn"], higher=True)
print("AFTER interaction filter:")
metrics("interaction contacts", lambda r:r["nkey"], higher=True)
metrics("gate+contacts+CNN combo", lambda r:combo(r), higher=True)
```

Supplementary S4. LLM citation-accuracy audit: the complete list of claim-citation pairs verified for each case, the accept/reject verdict and reason for each, the cited PMID and DOI, and the per-case and pooled Wilson confidence intervals. This is the underlying record for Table 2

LLM Citation-Accuracy Audit

Underlying record for Table 2. For each case, every claim-citation pair present in the program’s LLM-generated literature-synthesis / target-annotation outputs was verified directly against PubMed (all available pairs, not a random sample). ACCEPT = cited reference supports the claim; REJECT = irrelevant, contradictory, or the identifier does not resolve. Single non-blinded author audit; inter-rater

agreement not measured. Epidemiology/market-size citations are out of the target-biology/SAR/activity scope of this audit and are not included. Redaction notice: to protect confidential program information for which provisional patent protection has not yet been filed, the specific claim wording, cited PMIDs, and DOIs are withheld in this reviewer-facing copy. Each pair’s case assignment and accept/reject verdict are retained so that the per-case counts, accuracy, and confidence intervals in the summary table remain fully auditable. The complete per-pair record, including all PMIDs and DOIs, is available to the Editor and, through the Editor, to reviewers confidentially on request, and will be released publicly upon provisional filing.

Per-case summary

Case	n	Accept	Accuracy	Wilson 95% CI
A	8	7	88%	53-98%
B	9	7	78%	45-94%
C	10	9	90%	60-98%
Pooled	27	23	85%	not computed (clustered; descriptive only)



Full per-pair audit

Case	Claim (abbrev.)	PMID	Verdict	Reason	DOI
A	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
A	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
A	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
A	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	REJECT	Citation–claim mismatch (cited reference does not support the claim)	[withheld]
A	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
A	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
B	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
B	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
B	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
B	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
B	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	REJECT	Citation–claim mismatch (cited reference does not support the claim)	[withheld]
B	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	REJECT	Citation–claim mismatch (cited reference does not support the claim)	[withheld]

Continued

Case	Claim (abbrev.)	PMID	Verdict	Reason	DOI
B	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
B	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
B	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
C	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
C	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
C	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
C	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
C	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
C	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
A	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
A	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
C	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]

Continued					
Case	Claim (abbrev.)	PMID	Verdict	Reason	DOI
C	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
C	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	ACCEPT	Citation verifiably supports the claim on independent PubMed review	[withheld]
C	[Target-biology / SAR / activity claim — withheld pending provisional filing]	[withheld]	REJECT	Cited identifier does not resolve to a real PubMed record	[withheld]

Source: PubMed (article metadata). CIs computed with statsmodels 0.14.6 proportion_confint (method='wilson').