DORA: DISCOVERY-ORIENTED AGENT FOR DRUG REPURPOSING

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific discovery requires connecting fragmented evidence to construct and verify hypotheses that are not explicitly documented. We tackle this challenge through a case study on drug repurposing: identifying new therapeutic uses for existing drugs by connecting evidence across molecular targets, pathways, phenotypes, biological processes and disease mechanisms. We construct a comprehensive benchmark from approximately 34k clinical trials and 60k drug–disease pairs. We introduce DORA, a Discovery-Oriented Agent for drug repurposing, guided by an agentic environment-driven paradigm for scientific exploration. We construct a comprehensive biomedical knowledge network consisted of 273K nodes and 16.22M edges from a massive amount of heterogeneous databases, literature, and scientific tools, and design a novel algorithm, Meta-Pathfinder, to uncover non-trivial, implicit mechanistic paths connecting drugs to diseases. Rather than executing a predefined workflow, an LLM operates through a general-purpose coding-agent harness to interact dynamically with this environment: invoking Meta-Pathfinder and scientific tools, traversing evidence, identifying missing links, challenging competing hypotheses, and iteratively refining therapeutic hypotheses. Across seven frontier models, DORA improves five LLM-only baselines, with gains of 2.45–28.09 normalized-score points and 1.00–16.00 percentage points in categorical accuracy; two backbones do not improve. Notably, smaller models equipped with our agentic environment can outperform substantially larger models without it. To our knowledge, this is the first systematic framework for agentic drug repurposing, moving beyond retrieving and reasoning over existing knowledge to connecting fragmented evidence, uncovering latent mechanisms, and generating novel therapeutic hypotheses. We will publicly release our platform, data, and resources to support reproducible research.

1 INTRODUCTION

Scientific discovery does not always begin with a brand new observation; it can also emerge from a new connection: linking pieces of knowledge that are individually known yet have never been brought together to reveal a useful hypothesis. In this work, we focus on systematically discovering such connections for drug repurposing, which seeks new therapeutic indications for existing drugs, offering a potentially faster and lower-risk alternative to de novo development (Ashburn & Thor, 2004; Langedijk et al., 2015). This advantage is substantial given the high cost of developing new molecular entities. Historically, repurposing has emerged through either serendipitous observation or mechanism-driven hypothesis testing. Sildenafil, for example, was repurposed from angina to erectile dysfunction following unexpected clinical effects (Goldstein et al., 2019), while thalidomide was repurposed for multiple myeloma based on its antiangiogenic and immunomodulatory activity (Singhal et al., 1999). Mechanism-driven repurposing is particularly promising for rare diseases, where existing clinical characterization can reduce development cost and risk (Sardana et al., 2011). A striking example is lymphangioleiomyomatosis (LAM): linking TSC1/TSC2 loss to aberrant mTORC1 activation revealed an unexpected therapeutic connection to sirolimus, an existing mTOR inhibitor used as an immunosuppressant. The central challenge is to integrate fragmented observations spanning molecular targets, pathways, cellular states, disease mechanisms, and clinical evidence into testable drug–disease hypotheses, rigorously evaluate supporting and contradictory evidence, and determine whether these connections can translate into therapeutic benefit.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

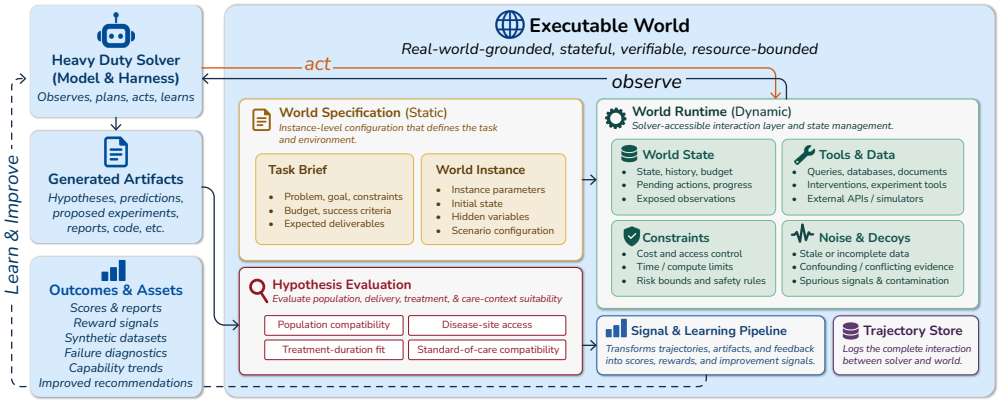

Figure 1: Overview of DORA and its environment-driven framework for drug repurposing.

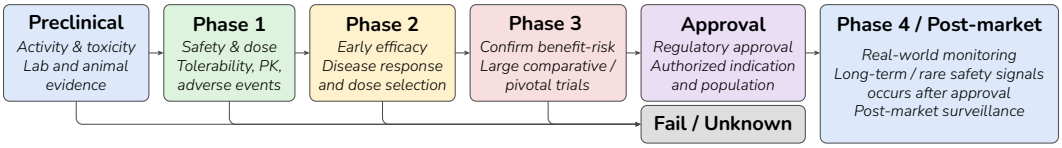

Figure 2: A FDA clinical trial needs to pass multiple phases until it final approval, but might also fail in any intermediate phases. We let systems predict how far the trial could potentially progress (more details in Section 2 and Table 2) and evaluate this capability on a frontier dataset.

Existing computational approaches largely frame drug repurposing as drug–disease link prediction, prioritizing candidates without fully interrogating their mechanistic evidence. LLM agents offer a complementary paradigm: integrating dispersed biomedical evidence, tracing mechanistic connections, and generating and evaluating repurposing hypotheses. We introduce **DORA**, an interactive agentic environment for drug repurposing (Figure 1). We construct an interconnected biomedical knowledge network integrating heterogeneous datasets, scientific literature, and tool-generated evidence, and introduce a Meta-pathfinder that searches this network for non-obvious, multi-hop mechanistic connections between drugs and disease processes. We connect this environment to an LLM agent through a general-purpose coding agent harness, which can invoke the Meta-pathfinder, inspect candidate mechanistic paths, gather supporting or contradictory evidence, perform targeted analyses, and iteratively refine its hypotheses. To evaluate our new paradigm, we construct a prospective benchmark from approximately 34k clinical trials and 60k drug–disease pairs. Across seven frontier models, DORA improves five LLM-only baselines by 2.45–28.09 normalized-score points and 1.00–16.00 categorical-accuracy percentage points. Smaller models with our environment can even outperform larger models without it, demonstrating the value of environment-driven discovery (Table 4). More interestingly, human experts identify novel mechanistic paths discovered by DORA that reveal previously unrecognized drug–disease connections (§ 4.3).

2 Task Definition

We formulate the drug repurposing task following Wang et al. (2026) as a forecasting task on trials. *Each trial is conducted on a pair of disease and drug.* We first associate clinical outcomes with ordered tiers according to Table 2. Let the evaluation set be $\mathcal{D} = \{(x_i, y_i, \tau_i)\}_{i=1}^{N}$, where x_i is a drug–disease pair, $y_i \in [0, 1]$ is its observed clinical-outcome tier, and $\tau_i \in \{\text{exact}, \text{lower bound}\}$. Exact labels represent terminal outcomes; lower bounds record a stage attained without asserting that the program will progress no further. The system is required to output a promisingness score $\hat{y} \in [0, 1]$ and an evidence-grounded rationale.

Scoring A trial that we evaluate on can have one of the two types of status. Case 1: It might have been *closed*, which means it has either been approved or been rejected in an earlier phase. Case 2 is anything otherwise: it might still be *open*, which means it has passed an early phase, and might or might not have started the progress for the next phase. We evaluate two cases of trials

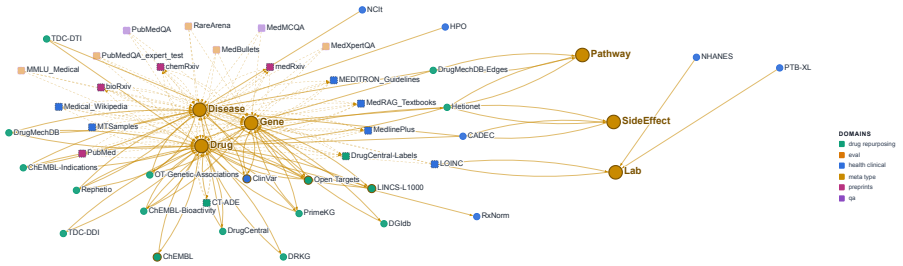

Figure 3: Agent-accessible datasets integrated into a heterogeneous knowledge graph. Large gold circles represent meta-types. Dataset nodes are depicted as circles for structured resources and squares for text, document, or benchmark resources. Solid edges indicate direct membership, while dashed edges denote membership inferred through entity extraction.

differently. For open trials, we define a loss as $l = \max(y_i - \hat{y}_i, 0)$. This takes into consideration the fact that the trial might move forward in the future. For closed trials, we define the loss as $|\hat{y}_i - y_i| + k \max(\hat{y}_i - y_i, 0)$. k equals the ratio between the number of open and closed trials, and counter-balances the tolerance to overscoring in open trials. Theoretically, when a trial is not known to be closed or open, the expected loss landscape is balanced and it is not advantageous to over- or under-predict the promisingness score. We define the closeness score as $q_i = 1 - \ell_i$, and the raw *continuous* prediction score is the arithmetic average of q_i among the evaluation dataset.

Normalization. To better interprete the prediction quality, we further renormalize the raw prediction score between the *optimal predictor performance* and *random guess performance*: $S_{\text{norm}} = \frac{S - S_R}{S_O - S_R} \times 100$. where $S_O = 1$ is achieved by an *optimal predictor* that knows the answer, and S_R is computed on a random guess predictor by drawing from the label distribution.

Categorical accuracy. For a complementary categorical measure, we also calculate the rate at which the predicted trial phase matches the ground truth. We categorize the predicted promisingness score by matching it to the closest anchor in Table 2. On open trials, the prediction is deemed as correct as long as the predicted clinical phase is higher than the current achieved phase.

3 METHOD

We introduce **DORA**, a discovery-oriented agent for drug repurposing. A general-purpose coding agent can attempt this task, but it must first work out which evidence matters, where to find it, and how to connect it, spending much of its budget on scaffolding rather than on the science. DORA moves this scaffolding into the environment (Figure 1): the agent starts with the evidence, connections, and tools that drug repurposing requires, and can focus on reasoning with them. The environment has two core components: an interconnected biomedical knowledge network built from heterogeneous databases, literature, and tool outputs (Section 3.1), and Meta-Pathfinder, a search algorithm that surfaces non-obvious mechanistic paths from a drug to a disease (Section 3.2). We treat such paths as candidate explanations, weighted by the strength and consistency of their evidence. The agent can also call simulation tools to generate evidence outside of existing databases to assist its predictions (Section 3.3). The agent runs in the `pi` coding-agent harness (Zechner, 2025), following the setup of Wang et al. (2026).

3.1 A HETEROGENEOUS KNOWLEDGE NETWORK

We assemble heterogeneous non-benchmark data sources spanning curated drug–disease and drug–target databases, molecular interaction and pathway graphs, pharmacovigilance resources, clinical and health text corpora, and external evaluation resources. *All data are date-gated to prevent data leakage of the newest trial outcomes used in evaluation.* Figure 3 illustrates how these sources are interconnected through shared biomedical entity types, while Table 1 summarizes their composition and scale. We expose each provider through a unified interface supporting three core operations: entity resolution, which maps a query entity (e.g., a drug) to source-specific nodes; neighborhood retrieval, which returns a bounded local subgraph; and evidence retrieval, which exposes detailed provenance for selected nodes and edges on demand. Importantly, each edge retains its source-

Table 1: Source groups of the agent-facing Biomedical KG. Scales follow dataset order; literature sources contribute metadata.

Group	Datasets	Scale (records unless noted)
Integrated KGs	DRKG, Hetionet, PrimeKG	97K / 47K / 129K nodes; 5.87M / 2.25M / 8.10M edges
Drug / target	ChEMBL, Open Targets, DrugMechDB, DrugCentral, LINCS-L1000, DGIdb, TDC-DTI, TDC-DDI	18,649 / 187,762 / 4,846 / 4,995 / 22.1M / 73,264 / 94,126 / 153,436
Clinical & terms	RxNorm, NCIt, HPO, LOINC, PTB-XL, MTSamples, NHANES	587K / 443K / 328K / 280K / 21,797 / 2,357; 12 cycles
Safety / AE	CT-ADE, CADEC	315K / 1,248
Health text	MedlinePlus, Medical Wikipedia, MedRAG Textbooks	1,016 / 36,800 / 125,764
Literature	PubMed, bioRxiv, medRxiv, chemRxiv	1.73M / 245K / 58K / 26K
Variant / pathway	ClinVar, gene–pathway graph	4.45M / 2.82M
QA & eval	MedMCQA, PubMedQA, MedBullets, MedXpertQA, MMLU-Medical, RareArena	187,005 / 273,018 / 596 / 2,450 / 1,087 / 72,661

specific score—such as literature confidence or interaction strength—while maintaining a separate search cost used for cross-provider pathfinding.

Structured biomedical knowledge. Datasets are connected through a shared schema of drugs, diseases, genes, proteins, pathways, variants, side effects, and phenotypes, enabling new sources to be integrated by simply mapping them to these types. Cross-source identifier mappings canonicalize entities using stable identifiers (e.g., ChEMBL/InChIKey for compounds, MONDO for diseases, and HGNC/Ensembl for genes), allowing records from different sources to resolve to the same node. Text fields are pre-indexed with dictionary-based entity recognition, enabling efficient entity linking in clinical notes and trial descriptions without query-time corpus scans.

Gene–pathway network. Many repurposing opportunities emerge only through mechanistic connections; for example, the sirolimus–LAM link in Section 1 becomes apparent through TSC1/TSC2 loss and mTORC1 signaling. We therefore construct a human gene–pathway network integrating NCBI Gene, BioGRID, Reactome, WikiPathways, and GO biological processes. After canonicalizing gene symbols, aliases, Entrez IDs, and Ensembl IDs, the network contains 197,674 gene nodes, 35,071 pathway and biological-process nodes, and 2.82M edges.

Literature Wiki/KG. We construct a literature provider from approximately 785K PubMed papers. We apply GLM-5.2 [1] to independently extract claims from each paper, which are then organized by entity and drug–disease hypothesis and synthesized into citation-grounded, wiki-style pages. The resulting graph contains 400,358 nodes (195K hypotheses, 87K targets, 73K drugs, 26K diseases, and 19K mechanisms), 195,501 directed drug–disease hypothesis edges, and 4.96M inter-page links. During exploration, the agent searches only this compact graph; for selected paths, the provider retrieves the underlying pages, citations, and contradictory evidence on demand. Appendix D details the construction pipeline.

As illustrated in Figure 4, candidate drug–disease connections can be traced through multi-step mechanistic paths spanning molecular, functional, phenotypic, and clinical layers. A path may link a drug to its molecular targets, genes, or signatures; propagate through pathways, biological processes, cell types, and cell states; and ultimately connect to disease phenotypes, symptoms, therapeutic effects, or adverse events. Individual links can be grounded in complementary evidence from scientific literature, experimental data, real-world data, pharmacovigilance, regulatory sources, and curated knowledge. Rather than treating these paths as established mechanisms, we use them as candidate mechanistic hypotheses whose plausibility is determined by the strength, consistency, and compatibility of the supporting evidence.

3.2 META-PATH FINDER

Single knowledge edges rarely connect distant and less studied pairs of diseases and drugs. To better model the usually weak relation between them, we take advantage of the comprehensive heterogeneous knowledge we collected above as connecting bridges. We propose Meta-Pathfinder, a graph-search algorithm. Given a drug and a disease, it searches for multi-edge connections across heterogeneous biomedical knowledge, including Drug, Disease, Gene, Pathway, SideEffect, and

[1] https://z.ai/blog/glm-5.2

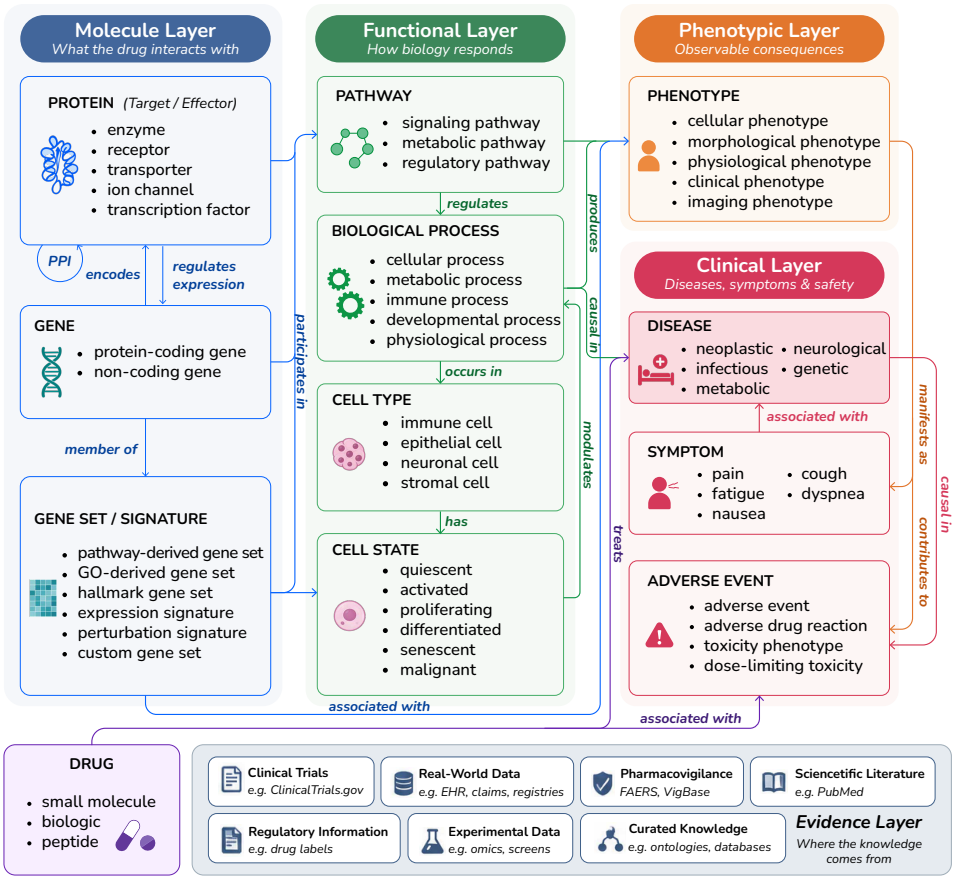

Figure 4: Interconnected biomedical knowledge network used to construct and evaluate mechanistic paths between drugs and diseases.

Labs illustrated in Figure 3. It performs $A^\star$ path search over the whole compatible graph, and merges the returned path skeletons while retaining source-specific identifiers and evidence. A returned path is treated as candidate evidence: it may expose a mechanistic chain, such as drug–target–pathway–disease, or a broader logical or analogical connection that points the agent to evidence worth examining, such connecting a drug X to a disease Y through its known cure X' that happens to target the same protein target T as X: $X - T - X' - Y$.

3.3 Simulation Tools

In addition, we equip the agent with NVIDIA's BioNeMo toolkit (NVIDIA), spanning structure prediction and co-folding, docking, molecular generation, and protein design. Rather than relying solely on existing knowledge, the agent can invoke these tools to computationally test emerging hypotheses—for example, evaluating a Meta-Pathfinder-proposed drug–target interaction through docking—and integrate the resulting predictions with retrieved evidence. In practice, however, the agent rarely calls these tools, relying mostly on the knowledge network and literature.

3.4 Drug Reformulation

To make our final report more informative, we extend repurposing from identifying what drug may work to hypothesizing *for whom and how* it should be deployed. Identifying a promising drug is only part of repurposing; its clinical utility may also depend on for whom and how it is deployed. After repurposing prediction, we further add a reformulation module that translates each drug–disease hypothesis into actionable development strategies along four dimensions: biomarkers, patient strat-

Table 2: Mapping trial phases to promisingness targets. An exact label specifies a single score; a lower bound admits any score from the floor to 1.00.

Trial phase	Associated value
Phase I failure or mixed	$\{0.00\}$
Phase I success; ongoing Phase II	$[0.15, 1.00]$
Phase II failure or mixed	$\{0.15\}$
Phase II success; ongoing Phase III	$[0.45, 1.00]$
Phase III failure or mixed	$\{0.45\}$
Phase III success; ongoing Phase IV	$[0.65, 1.00]$
Phase IV failure or mixed	$\{0.65\}$
Phase IV success or approval	$\{1.00\}$

Table 3: Dataset statistics for the public development set and trial corpus. Pair labels are those of each pair's most favorable record.

Quantity	Corpus	Dev Set
Drug–disease pairs	60,293	100
with post-cutoff approval	13	4
with exact target	159	30
Registry trials	33,634	124
Success	2,866	19
Failure	4,422	24
Mixed	679	4
Ongoing	25,667	77

ification, formulation and delivery, and combination therapy. Appendix E provides implementation details and an automatic evaluation when laboratory validation is unavailable.

3.5 Hypothesis Self-Reevaluation

After a candidate drug–disease pair is identified and its mechanism of action is established, we perform a second-stage hypothesis evaluation to assess whether the drug is clinically suitable for the new indication. The agent searches regulatory sources, including FDA prescribing information (U.S. Food and Drug Administration, 2026), DailyMed (U.S. National Library of Medicine, 2026), and relevant literature, to evaluate population compatibility, disease-compartment access, treatment-duration suitability, and compatibility with standard-of-care therapies (Figure 1). In particular, when relevant to the target population, it assesses suitability for pediatric, elderly, pregnant, or renally/hepatically impaired patients; whether the drug can reach the relevant disease site, such as crossing the blood–brain barrier, penetrating ocular tissue, achieving sufficient lung exposure, or entering intracellular compartments; whether its safety profile is compatible with the required treatment duration; and whether it can be used alongside existing therapies for the disease.

4 Experiments

4.1 Data

We construct a prospective drug-repurposing benchmark from interventional clinical trials in ClinicalTrials.gov (Zarin et al., 2011) and the EU Clinical Trials Information System (European Medicines Agency). Mapping each trial to its tested drugs and target diseases yields 60,293 drug–disease pairs across 33,634 trials. To reduce outcome leakage, we retain only pairs whose most recent registry evidence is dated on or after January 1, 2026 (Appendix B.3). To better reflect realistic drug repurposing scenarios, we further require that each drug is an existing compound and was not approved for the target disease before the cutoff date. This yields 547 eligible pairs across 784 trials. Because each evaluation requires a full agent run, we randomly sample 100 pairs among these to form a public development set. Appendix B provides further details regarding constructions of both benchmarks, and Table 3 summarizes the dataset statistics. An LLM judge extracts each trial's outcome from its registry record, and a human expert reviews and corrects every label. For pairs with multiple trials, we use the most favorable trial-level label. Endpoint interpretation, terminated trial handling, and other edge cases are detailed in the Appendix.

4.2 Experiment Results

Table 4 shows that DORA improves five of seven backbones, with the largest gain for Qwen3.8 Max (+28.09 normalized-score points and +16.00 categorical-accuracy percentage points). Kimi K3 and GPT-5.6-sol gain 9.53 and 7.60 normalized-score points, respectively, while DeepSeek V4 Pro and GLM-5.2 do not improve. The heterogeneous response shows that baseline model strength alone does not determine the benefit of the environment. We also compare our DORA environment with Biomni (Huang et al., 2025a), a widely used biomedical AI agent that integrates domain-

Table 4: Outcome-prediction performance on the 100-pair public development set. Evaluation is defined in Section 2 Values are mean ± sample standard deviation over three independent runs.

Backbone	Condition	Normalized continuous score	Raw continuous score	Categorical accuracy
Claude Opus 5	LLM-only	64.18 ± 1.09	0.9532 ± 0.0014	$(87.67 \pm 0.58)\%$
	DORA	66.63 ± 1.17	0.9564 ± 0.0015	$(88.67 \pm 0.58)\%$
GPT-5.6-sol	LLM-only	54.21 ± 2.20	0.9401 ± 0.0029	$(80.67 \pm 1.15)\%$
	DORA	61.81 ± 3.92	0.9501 ± 0.0051	$(85.00 \pm 1.73)\%$
GPT-5.5	LLM-only	53.78 ± 3.22	0.9396 ± 0.0042	$(81.00 \pm 1.73)\%$
	DORA	56.30 ± 1.56	0.9429 ± 0.0020	$(83.67 \pm 1.15)\%$
Kimi K3	LLM-only	44.70 ± 4.63	0.9277 ± 0.0061	$(75.00 \pm 2.65)\%$
	DORA	54.24 ± 3.10	0.9402 ± 0.0041	$(82.00 \pm 1.73)\%$
Qwen3.8 Max	LLM-only	24.05 ± 3.50	0.9007 ± 0.0046	$(64.00 \pm 2.65)\%$
	DORA	52.15 ± 2.98	0.9374 ± 0.0039	$(80.00 \pm 2.00)\%$
DeepSeek V4 Pro	LLM-only	44.68 ± 2.56	0.9277 ± 0.0034	$(76.67 \pm 1.53)\%$
	DORA	39.22 ± 0.77	0.9205 ± 0.0010	$(76.00 \pm 3.00)\%$
GLM-5.2	LLM-only	37.74 ± 4.74	0.9186 ± 0.0062	$(74.67 \pm 1.15)\%$
	DORA	36.51 ± 4.86	0.9170 ± 0.0064	$(74.25 \pm 3.03)\%$

Table 5: Comparison with Biomni-resource configurations. Prediction score refers to the normalized continuous score across three seeds and C denotes confidence calibration.

Configuration	Prediction score	C
Biomni: static evidence	57.27 ± 1.30	0.7299
Biomni: callable tools	54.65 ± 1.66	0.7282
DORA	$\mathbf{61.81 \pm 3.92}$	**0.8293**

specific databases and analysis tools. We consider two configurations: provide Biomni access to drug records, biomedical knowledge graphs, and drug–disease predictions through either *static evidence*, which supplies precomputed results in the prompt, or *callable tools*, which allows the agent to query these resources during inference. Both use bounded adapters rather than the unrestricted native Biomni agent to avoid ground truth leakage same as our main experimental setup. All configurations use GPT-5.6-sol with medium reasoning and are evaluated on the same 100 drug–disease pairs over three stochastic replicates.

Ablation Study Drug-repurposing prediction is frontier and intrinsically challenging, with model prediction demonstrating a large variance given the small evaluation set and tight computation budget. Consequently, isolated point-to-point ablations can introduce a large uncertainty in the trends. We therefore use a *combinatorial ablation* design. We evaluate multiple component combinations (Appendix Section C) and infer their marginal linear contributions jointly, rather than interpreting every pairwise difference as an independent effect. Table 6 reports the inferred effect β_k for the components evaluated in the main paper. A positive coefficient favors retaining or adding the component. Meta-Pathfinder retains a positive 95% credible interval under the adjusted model. Biomedical KG has a negative interval in this additive fit, while the Literature Wiki and BioNeMo intervals cross zero.

4.3 Analysis

Pathfinding Qualitative Analysis Meta-Pathfinder makes intermediate graph connections available for review. Figure 5 shows three example routes retrieved from evaluation results. In the first example, dupilumab treats eczema by blocking two immune signals that drive allergic inflammation. Eczema is strongly linked to FLG, a gene that helps maintain the skin barrier; when that barrier is weakened, allergens such as pollen can enter more easily and train the immune system to react to them. That same allergic response can later appear in the nose as seasonal allergic rhinitis. The path connects two seemingly disconnected diseases through a shared allergic mechanism, suggesting that a drug that dampens this inflammation in the skin may do the same in the nose. In the second one, pembrolizumab treats classic Hodgkin lymphoma by blocking the PD-1/PD-L1 pathway, which tumors use to shut down attacks from the body's own immune cells. The gene CD274

Table 6: Ablation estimates for component contribution to the normalized scores. Positive β favors retaining or adding a component, and $p_+ = P(\beta_k \leq 0) = 1 - P(\beta_k > 0)$ is the prespecified one-sided Bayesian tail probability for measuring the statistical significance.

Component present	Mean $\beta \uparrow$	95% credible interval $\uparrow$	$P(\beta > 0) \uparrow$	$p_+ \downarrow$
Literature Wiki	$+0.67$	$[-4.06, +5.39]$	0.609	0.391
BioNeMo	$+2.54$	$[-1.99, +7.07]$	0.864	0.136
Biomedical KG	-4.67	$[-7.94, -1.39]$	0.003	0.997
Meta-Pathfinder	$+4.19$	$[+1.11, +7.27]$	0.996	0.004

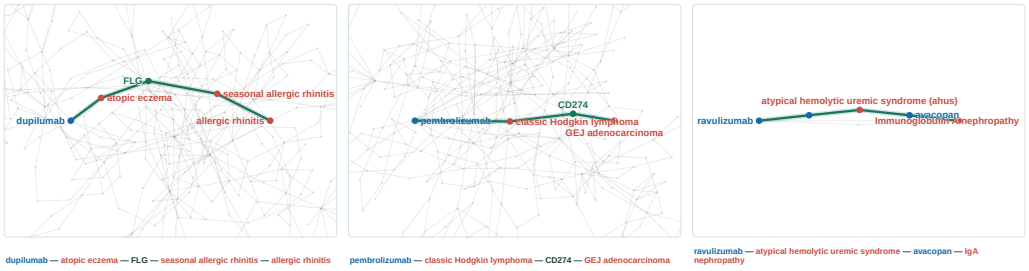

Figure 5: Candidate connection paths retrieved by Meta-Pathfinder from saved run-local review graphs. Colored nodes and edges mark a selected route; the gray background preserves its local graph context.

produces the PD-L1 protein and is especially active in classic Hodgkin lymphoma, helping explain why pembrolizumab works well there. PD-L1 is also active in gastroesophageal junction adenocarcinoma, where tumors can use the same mechanism to escape the immune system. Although the two cancers arise in different tissues, the path shows that both rely on the same immune-escape mechanism, which offers a direct rationale for extending pembrolizumab to the second. In the third example, ravulizumab treats atypical hemolytic uremic syndrome by blocking C5, part of an immune pathway called the complement system that can damage blood vessels when overactive. Avacopan blocks another part of this same pathway, showing that both drugs target closely related biology. In IgA nephropathy, abnormal immune deposits in the kidney can also activate the complement system and contribute to kidney damage. This path links two very different diseases through the same set of pathways, giving a potential reason why a complement-blocking drug might be useful in both.

Remaining Error analysis. The remaining errors fall into two main failure modes. First, the agent reasons from biological and clinical evidence but cannot anticipate non-scientific factors, such as trial sponsor-level decisions, that affect trial progression. For riociguat in early pulmonary vascular disease, a precursor to pulmonary arterial hypertension, the agent predicted 0.82 against a target of 0.15 (failed Phase II). The trial met its primary endpoint, consistent with the agent's reasoning, but was terminated early after the drug supplier withdrew for commercial reasons, and the program did not advance. Second, the agent can overgeneralize evidence from a disease subgroup to the broader indication. For nivolumab in metastatic colorectal cancer, it predicted 1.00 against a target of 0.15 (failed Phase II). Nivolumab is effective in the small subset of colorectal tumors with high mutational burden, but the agent generalized this evidence to colorectal cancer overall. Most colorectal tumors lack this feature and respond poorly to immunotherapy; accordingly, adding nivolumab to chemotherapy did not improve progression-free survival in the trial.

5 Related Work

5.1 AI for Drug Repurposing

Previous graph-based discovery methods (Himmelstein et al., 2017b; Luo et al., 2017; Zeng et al., 2019; Cui et al., 2021; Wang et al., 2021; Chandak et al., 2023; Ma et al., 2023; Huang et al., 2024). are bounded by encoded relations and fixed schemas, limiting its ability to pursue hypotheses that

require missing links. LLMs offer more flexible reasoning (Xiang et al., 2025; Saranya et al., 2026; Sanz et al., 2025), but their parametric knowledge can be incomplete and outdated. Recent agents begin to bridge this gap: DrugAgent integrates heterogeneous evidence for drug–target assessment (Inoue et al., 2024), while BioDiscoveryAgent combines tools with experimental feedback (Roohani et al., 2025). Yet generating plausible hypotheses is not sufficient: they must be grounded in evidence and mechanistically coherent (Pushpakom et al., 2019; Singh, 2024; Wan et al., 2025). We therefore construct an executable environment to discover mechanistic paths, inspect supporting and conflicting evidence, and iteratively refine them through further analysis.

5.2 Knowledge-enhanced Agentic AI

Recent knowledge-enhanced agents move beyond fixed retrieval toward agent-controlled knowledge acquisition. Self-RAG learns when to retrieve and critique evidence (Asai et al., 2024), while PaperQA2 iteratively searches, synthesizes, and evaluates scientific literature (Skarlinski et al., 2024). Structured environments further enable active, auditable exploration: Think-on-Graph performs iterative multi-hop graph reasoning (Sun et al., 2024), GeneGPT queries curated genomic APIs (Jin et al., 2024), and TxAgent dynamically orchestrates 211 biomedical tools (Gao et al., 2025). Together, these systems establish agents as controllers that actively seek and integrate external knowledge rather than passively consuming retrieved context. Scientific agents extend this paradigm from knowledge acquisition to hypothesis-driven discovery. ChemCrow (M. Bran et al., 2024) and Coscientist (Boiko et al., 2023) orchestrate tools for chemical reasoning and experimentation; MOOSE-Chem (Yang et al., 2025) and BioDiscoveryAgent (Roohani et al., 2025) iteratively generate and refine scientific hypotheses. In pharmaceutical development, DrugAgent integrates predictive models, knowledge graphs, and literature for repurposing assessment (Inoue et al., 2024), while APPA combines heterogeneous evidence and models for preformulation planning (Lange et al., 2025).

5.3 Evidence Finding from Knowledge Graphs

Path-based reasoning is widely used for knowledge-graph inference. DeepPath (Xiong et al., 2017) learns policy-guided walks, PoLo (Liu et al., 2021) incorporates logical rules into interpretable biomedical paths, and K-Paths (Abdullahi et al., 2025) retrieves diverse loopless paths as compact inputs to downstream models. Despite different objectives, these methods typically operate over a fixed, pre-integrated graph and use its existing topology to recover relations or answer queries. For drug repurposing, Rephetio (Himmelstein et al., 2017a) ranks drug–disease candidates using typed meta-paths, while KGML-xDTD (Ma et al., 2022) and XG4Repo (Jiménez et al., 2024) combine relation prediction with mechanistic paths; DrugMechDB (Gonzalez-Cavazos et al., 2023) instead curates mechanism paths for known indications. These approaches largely optimize relation prediction or path recovery over an integrated graph. In contrast, our Meta-Pathfinder searches across independent knowledge providers connected through shared entity types while preserving source-local provenance. Rather than treating a retrieved path as an answer, we expose it as a candidate mechanistic hypothesis that the agent can interrogate for supporting evidence, contradictions, and missing links.

6 Conclusions and Future Work

We introduced DORA, a discovery-oriented agent for drug repurposing that enables LLMs to interact with an executable environment integrating heterogeneous biomedical knowledge, literature, and scientific tools. DORA discovers mechanistic paths, evaluates supporting and conflicting evidence, and iteratively refines therapeutic hypotheses. Across seven frontier models, DORA consistently improves over LLM-only baselines, with smaller environment-equipped models sometimes outperforming substantially larger models without the environment. These results suggest that the knowledge-rich environment itself can be a key scaling axis for scientific discovery. Future work will incorporate richer experimental and multimodal evidence, improve reasoning over uncertainty and competing hypotheses, and close the loop with prospective experimental validation. More broadly, we hope this work motivates a shift from scaling models that reason over existing knowledge toward building environments in which agents can actively explore, test, and discover.

AI DISCLOSURE

In this work, we used generative AI tools to implement some methods and interpret some results. We have not used generative AI tools for designing or providing feedback on research methodology or experiments, and generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refine hypotheses are not applicable to this work. Additionally, we used generative AI tools for editing some sections to improve readability. We have reviewed all AI-assisted work. Human experts checked the qualitative results from the experiments and confirmed its validity. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Tassallah Abdullahi, Ioanna Gemou, Nihal V. Nayak, Ghulam Murtaza, Stephen H. Bach, Carsten Eickhoff, and Ritambhara Singh. K-Paths: Reasoning over graph paths for drug repurposing and drug interaction prediction. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 5–16, 2025. doi:10.1145/3711896.3737011.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

Ted T. Ashburn and Karl B. Thor. Drug repositioning: identifying and developing new uses for existing drugs. *Nature Reviews Drug Discovery*, 3(8):673–683, 2004. doi:10.1038/nrd1468.

Sorin Avram et al. DrugCentral 2023 extends human clinical data and integrates veterinary drugs. *Nucleic Acids Research*, 51(D1):D1276–D1287, 2023. doi:10.1093/nar/gkac1085.

Daniil A Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023.

Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific data*, 10(1):67, 2023.

Code of Federal Regulations. Title 21, section 312.85: Phase 4 studies. https://www.ecfr.gov/current/title-21/chapter-I/subchapter-D/part-312/subpart-B/section-312.85. Accessed September 2026.

Chen Cui, Xiaoyu Ding, Dingyan Wang, Lifan Chen, Fu Xiao, Tingyang Xu, Mingyue Zheng, Xiaomin Luo, Hualiang Jiang, and Kaixian Chen. Drug repurposing against breast cancer by integrating drug-exposure expression profiles and drug–drug links based on graph neural network. *Bioinformatics*, 37(18):2930–2937, 2021.

European Medicines Agency. Clinical trials information system (CTIS): Public portal. https://euclinicaltrials.eu. Accessed September 2026.

Shanghua Gao, Richard Zhu, Zhenglun Kong, Ayush Noori, Xiaorui Su, Curtis Ginder, Theodoros Tsiligkaridis, and Marinka Zitnik. Txagent: an ai agent for therapeutic reasoning across a universe of tools. *arXiv preprint arXiv:2503.10970*, 2025.

Irwin Goldstein, Arthur L. Burnett, Raymond C. Rosen, Peter W. Park, and Vera J. Stecher. The serendipitous story of sildenafil: An unexpected oral therapy for erectile dysfunction. *Sexual Medicine Reviews*, 7(1):115–128, 2019. doi:10.1016/j.sxmr.2018.06.005.

Adriana Carolina Gonzalez-Cavazos, Anna Tanska, Michael Mayers, Denise Carvalho-Silva, Brindha Sridharan, Patrick A. Rewers, Umasri Sankarlal, Lakshmanan Jagannathan, and Andrew I. Su. DrugMechDB: A curated database of drug mechanisms. *Scientific Data*, 10:632, 2023. doi:10.1038/s41597-023-02534-z.

Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L. Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E. Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726, 2017a. doi:10.7554/eLife.26726.

Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, and Sergio E Baranzini. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *elife*, 6:e26726, 2017b.

Kexin Huang, Payal Chandak, Qianwen Wang, Shreyas Havaldar, Akhil Vaid, Jure Leskovec, Girish N Nadkarni, Benjamin S Glicksberg, Nils Gehlenborg, and Marinka Zitnik. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, 30(12):3601–3613, 2024.

Kexin Huang, Serena Zhang, Hanchen Wang, Yuanhao Qu, Yingzhou Lu, Yusuf Roohani, Ryan Li, Lin Qiu, Junze Zhang, Yin Di, et al. Biomni: A general-purpose biomedical ai agent. *bioRxiv*, pp. 2025–05, 2025a.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025b. ISSN 1558-2868. doi:10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

Yoshitaka Inoue, Tianci Song, and Tianfan Fu. Drugagent: Explainable drug repurposing agent with large language model-based reasoning. *arXiv preprint arXiv:2408.13378*, 2024. URL https://arxiv.org/abs/2408.13378v1. Version 1; accessed September 2026.

Ana Jiménez, María José Merino, Juan Parras, and Santiago Zazo. Explainable drug repurposing via path based knowledge graph completion. *Scientific Reports*, 14:16587, 2024. doi:10.1038/s41598-024-67163-x.

Qiao Jin, Yifan Yang, Qingyu Chen, and Zhiyong Lu. Genegpt: augmenting large language models with domain tools for improved access to biomedical information. *Bioinformatics*, 40(2):btae075, 2024.

Sunghwan Kim et al. PubChem 2025 update. *Nucleic Acids Research*, 53(D1):D1516–D1525, 2025. doi:10.1093/nar/gkae1059.

Craig Knox et al. DrugBank 6.0: The DrugBank knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275, 2024. doi:10.1093/nar/gkad976.

Julius Lange, Leonid Komissarov, Nicole Wyttenbach, and Andrea Anelli. Appa: agentic preformulation pathway assistant. *arXiv preprint arXiv:2503.16698*, 2025.

Joyce Langedijk, Aukje K. Mantel-Teeuwisse, Diederick S. Slijkerman, and Marie-Hélène D. B. Schutjens. Drug repositioning and repurposing: terminology and definitions in literature. *Drug Discovery Today*, 20(8):1027–1034, 2015. doi:10.1016/j.drudis.2015.05.001.

Yushan Liu, Marcel Hildebrandt, Mitchell Joblin, Martin Ringsquandl, Rime Raissouni, and Volker Tresp. Neural multi-hop reasoning with logical rules on biomedical knowledge graphs. In *The Semantic Web*, pp. 375–391. Springer International Publishing, 2021. doi:10.1007/978-3-030-77385-4_22.

Yunan Luo, Xinbin Zhao, Jingtian Zhou, Jinglin Yang, Yanqing Zhang, Wenhua Kuang, Jian Peng, Ligong Chen, and Jianyang Zeng. A network integration approach for drug-target interaction prediction and computational drug repositioning from heterogeneous information. *Nature communications*, 8(1):573, 2017.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D White, and Philippe Schwaller. Augmenting large language models with chemistry tools. *Nature machine intelligence*, 6(5):525–535, 2024.

Chunyu Ma, Zhihan Zhou, Han Liu, and David Koslicki. KGML-xDTD: A knowledge graph-based machine learning framework for drug treatment prediction and mechanism description. *GigaScience*, 12, 2022. doi:10.1093/gigascience/giad057.

Chunyu Ma, Zhihan Zhou, Han Liu, and David Koslicki. Kgml-xdtd: a knowledge graph–based machine learning framework for drug treatment prediction and mechanism description. *GigaScience*, 12:giad057, 01 2023. ISSN 2047-217X. doi:10.1093/gigascience/giad057. URL https://doi.org/10.1093/gigascience/giad057.

Adnan Malik et al. ChEBI: Re-engineered for a sustainable future. *Nucleic Acids Research*, 54(D1): D1768–D1778, 2026. doi:10.1093/nar/gkaf1271.

James Malone et al. Modeling sample variables with an Experimental Factor Ontology. *Bioinformatics*, 26(8):1112–1118, 2010. doi:10.1093/bioinformatics/btq099.

NVIDIA. Overview of BioNeMo. BioNeMo Framework documentation. URL https://docs.nvidia.com/bionemo-framework/latest/main/about/overview/. Accessed: 2026-09-20.

Sudeep Pushpakom, Francesco Iorio, Patrick A Eyers, K Jane Escott, Shirley Hopper, Andrew Wells, Andrew Doig, Tim Guilliams, Joanna Latimer, Christine McNamee, et al. Drug repurposing: progress, challenges and recommendations. *Nature reviews Drug discovery*, 18(1):41–58, 2019.

Yusuf Roohani, Andrew Lee, Qian Huang, Jian Vora, Zachary Steinhart, Kexin Huang, Alexander Marson, Percy Liang, and Jure Leskovec. Biodiscoveryagent: An ai agent for designing genetic perturbation experiments. In *International Conference on Learning Representations*, volume 2025, pp. 26417–26466, 2025.

Iratxe Zunzunegui Sanz, Belen Otero-Carrasco, and Alejandro Rodriguez-Gonzalez. Accelerating drug repurposing with ai: The role of large language models in hypothesis validation. In *2025 IEEE 38th International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 147–153. IEEE, 2025.

K. Saranya, Emerson Raja Joseph, Ts. Kalaiarasi, and M. Karthiga. Generative ai in drug repurposing and biomarker discovery: a multimodal approach. *Frontiers in Bioinformatics*, Volume 6 - 2026, 2026. ISSN 2673-7647. doi:10.3389/fbinf.2026.1755412. URL https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2026.1755412.

Divya Sardana, Cheng Zhu, Minlu Zhang, Ranga C. Gudivada, Lun Yang, and Anil G. Jegga. Drug repositioning for orphan diseases. *Briefings in Bioinformatics*, 12(4):346–356, 2011. doi:10.1093/bib/bbr021.

Anuradha Singh. Artificial intelligence for drug repurposing against infectious diseases. *Artificial Intelligence Chemistry*, 2(2):100071, 2024. ISSN 2949-7477. doi:https://doi.org/10.1016/j.aichem.2024.100071. URL https://www.sciencedirect.com/science/article/pii/S2949747724000290.

Seema Singhal, Jayesh Mehta, Raman Desikan, Dan Ayers, Paula Roberson, Paul Eddlemon, Nikhil Munshi, Elias Anaissie, Carla Wilson, Madhav Dhodapkar, Jerome Zeldis, David Siegel, John Crowley, and Bart Barlogie. Antitumor activity of thalidomide in refractory multiple myeloma. *New England Journal of Medicine*, 341(21):1565–1571, 1999. doi:10.1056/NEJM199911183412102.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *International Conference on Learning Representations*, volume 2024, pp. 3868–3898, 2024.

U.S. Food and Drug Administration. Drugs@FDA: FDA-Approved Drugs. https://www.fda.gov/drugs/drug-approvals-and-databases/drugsfda-data-files, 2026. Accessed: 2026-09-25.

U.S. National Library of Medicine. DailyMed. https://dailymed.nlm.nih.gov/dailymed/, 2026. Accessed: 2026-09-25.

Nicole A. Vasilevsky et al. Mondo: Unifying diseases for the world, by the world. *medRxiv*, 2022. doi:10.1101/2022.04.13.22273750.

Zhaoman Wan, Xinran Sun, Yi Li, Tianyao Chu, Xueyu Hao, Yang Cao, and Peng Zhang. Applications of artificial intelligence in drug repurposing. *Advanced Science*, 12(14):2411325, 2025.

Brian Wang, Bin Feng, Xiaoman Pan, Chenyang An, Felix Liu, Tangqi Fang, Gongbo Sun, Lingfeng Shen, Ning Wang, Handuo Zhang, Feng Chen, Fuchao Yang, Xiang Wang, Jiacheng Lin, Siting Li, Zixuan Liu, Chi Han, Zhenhailong Wang, Kunlun Zhu, Lawrence Zhao, Yueqi Guo, Kailong Wen, Feng Xing, Yiling Guo, Lidong Bing, David Tan, Bo An, Heng Ji, and Sheng Wang. Apodex discovery: Reality benchmarks and environments for evaluating and building discoverative artificial intelligence, 2026. URL https://arxiv.org/abs/2608.11341.

Qingyun Wang, Manling Li, Xuan Wang, Nikolaus Parulian, Guangxing Han, Jiawei Ma, Jingxuan Tu, Ying Lin, Ranran Haoran Zhang, Weili Liu, Aabhas Chauhan, Yingjun Guan, Bangzheng Li, Ruisong Li, Xiangchen Song, Yi Fung, Heng Ji, Jiawei Han, Shih-Fu Chang, James Pustejovsky, Jasmine Rah, David Liem, Ahmed ELsayed, Martha Palmer, Clare Voss, Cynthia Schneider, and Boyan Onyshkevych. COVID-19 literature knowledge graph construction and drug repurposing report generation. In Avi Sil and Xi Victoria Lin (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pp. 66–77, Online, June 2021. Association for Computational Linguistics. doi:10.18653/v1/2021.naacl-demos.8. URL https://aclanthology.org/2021.naacl-demos.8/.

Chengrui Xiang, Tengfei Ma, Xiangzheng Fu, Yiping Liu, Bosheng Song, and Xiangxiang Zeng. From knowledge to treatment: Large language model assisted biomedical concept representation for drug repurposing. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13967–13982, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi:10.18653/v1/2025.findings-emnlp.751. URL https://aclanthology.org/2025.findings-emnlp.751/.

Wenhan Xiong, Thien Hoang, and William Yang Wang. DeepPath: A reinforcement learning method for knowledge graph reasoning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 564–573, 2017. doi:10.18653/v1/D17-1060.

Zonglin Yang, Wanhao Liu, Ben Gao, Tong Xie, Yuqiang Li, Wanli Ouyang, Soujanya Poria, Erik Cambria, and Dongzhan Zhou. Moose-chem: Large language models for rediscovering unseen chemistry scientific hypotheses. In *International Conference on Learning Representations*, volume 2025, pp. 33251–33277, 2025.

Deborah A. Zarin, Tony Tse, Rebecca J. Williams, Robert M. Califf, and Nicholas C. Ide. The ClinicalTrials.gov results database — update and key issues. *New England Journal of Medicine*, 364(9):852–860, 2011. doi:10.1056/NEJMsa1012065.

Barbara Zdrazil et al. The ChEMBL database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1):D1180–D1192, 2024. doi:10.1093/nar/gkad1004.

Mario Zechner. Pi Agent Harness. GitHub repository, 2025. URL https://github.com/earendil-works/pi. Accessed: 2026-09-20.

Xiangxiang Zeng, Siyi Zhu, Xiangrong Liu, Yadi Zhou, Ruth Nussinov, and Feixiong Cheng. deepdr: a network-based deep learning approach to in silico drug repositioning. *Bioinformatics*, 35(24):5191–5198, 2019.

Kunlun Zhu, Zijia Liu, Bingxuan Li, Muxin Tian, Yingxuan Yang, Jiaxun Zhang, Pengrui Han, Qipeng Xie, Fuyang Cui, Weijia Zhang, Xiaoteng Ma, Xiaodong Yu, Gowtham Ramesh, Jialian Wu, Zicheng Liu, Pan Lu, James Zou, and Jiaxuan You. Where llm agents fail and how they can learn from failures, 2025. URL https://arxiv.org/abs/2509.25370.

A Verification Layer

A.1 Verification

Language models produce fluent statements that are not true (Huang et al., 2025b), and in a long-horizon agentic run an early unchecked error propagates into the decisions that follow it (Zhu et al., 2025). A repurposing trajectory is long and its conclusion rests on a chain of intermediate claims, so a well-written report is no evidence that the investigation behind it was sound. Verification therefore evaluates both the evidence supporting a prediction and the process that produced it. *Evidence verification* audits the claims an agent submits and tests whether they support the proposed indication. *Process verification* evaluates how the agent searched, observed, and revised. Neither has access to the held-out outcome, and neither changes the outcome score.

Evidence verification. Evidence verification operates on the mechanism paths and supporting claims an agent submits with its prediction. Each asserted link carries a cited record and a supporting span. We first check attribution: the source must state the relation claimed for it and must predate the evidence cutoff. Accurate attribution does not establish biological validity, so we re-examine the load-bearing links using data and edge attributes that Meta-Pathfinder (Section 3.2) ignores during traversal. The mechanism-direction check illustrates the pattern: it compares the drug's action on a target with the way that target changes in the disease, and a drug that pushes the target further in the disease direction counts against the hypothesis.

Biological and safety checks. The remaining checks are deliberate attempts at refutation. Confirming a link against the source that proposed it is not an independent test; a check is informative only if it could have contradicted the agent. Together they enforce the conditions a domain expert applies before taking a candidate seriously, and that a fluent mechanistic account can still violate: the target must be present in the affected tissue, and independent genetic and pathway evidence must connect it to the disease. Safety acts as a constraint rather than a score, since toxicity in the treated organ cannot be offset by a strong mechanism. Weaker signals adjust the score instead of deciding it. Appendix A gives the catalog, the sources, and the decision rules.

Process verification. An agent can assemble accurate citations after the fact without having run the corresponding search, so process verification inspects the full trajectory: tool calls, the observations returned, and the reasoning in between. An automated evaluator, blind to the outcome, scores the six capabilities of HDS6 (Wang et al., 2026) on a 26-item rubric, examining whether conclusions follow from what the tools returned and whether the agent revised its hypothesis when the retrieved evidence contradicted it. Trajectories with fabricated tool output are excluded, and capabilities a trajectory never exercised are left unrated rather than counted as failures. Table 13 reports these scores.

A.2 Edge-verifier catalog

Each edge type in the mechanism graph defines a class of mechanistic claims checked by a verifier. Table 7 lists the catalog. Checks use a calculation or simulator whenever one exists. A constrained model rubric is used only when no calculation applies, and rubric-scored edges still require retrieved evidence and a citation. Edge types with the same checking method may share a verifier.

A.3 Independence and reconciliation

A verifier must be able to disagree with the system on evidential grounds, rather than through noise. It must compute a quantity different from traversal, read at least one source, tool, or edge field unused by search, and be capable of returning a verdict opposite to the system's conclusion on some

Table 7: Edge-verifier catalog. "Method" is the quantity computed, which in every case differs from the graph traversal used during search. "Independent source" names a source, tool, or edge field that the search providers do not use.

Verifier	Edge checked	Method	Independent source
Network proximity	protein–protein	shortest-path proximity against a degree-preserving permutation null	STRING (search uses BioGRID)
Signature reversal	drug → gene expression	rank-based connectivity of a drug profile against a disease signature	LINCS release disjoint from the indexed one
Binding affinity	drug ⊣ protein	measured K_i/IC_{50} thresholds; docking score against decoy null	BindingDB, IUPHAR; structure predictors
Mechanism direction	directed regulation	association-weighted vote comparing the drug's action type with the target's direction of change in disease	Open Targets, ChEMBL; action-type field unread by search
Genetic support	gene → disease	best target–disease association blended with target coverage; extensible to colocalization and Mendelian randomization	Open Targets Genetics
Pathway coherence	protein/gene → pathway	overlap significance between drug-target and disease-gene pathway sets	Reactome
Gene-set enrichment	gene → gene set	hypergeometric or rank-based enrichment with multiplicity control	MSigDB, GO
Target expression	target → tissue	expression level of each target in the organ affected by the disease; absence refutes the route	Human Protein Atlas
Entity grounding	node identity	cross-service normalization to a single identifier	NCATS NameRes/NodeNorm
Toxicity veto	drug → organ	organ collision between drug toxicity and disease site; caps the aggregate	curated FDA toxicity reference sets (DILIrank, DILIst, DIRIL, DICTrank), PubChem toxicology
Overdose margin	drug dose axis	parsed label overdose severity	openFDA labels
Pharmacovigilance	drug → adverse event	disproportionality (reporting odds ratio and related statistics)	openFDA FAERS
Cell state / process	abstract process edges	single-cell signature enrichment; constrained rubric as fallback	cell atlases; GO biological process

input. Reusing an underlying database is acceptable when the check uses information that search ignores. For example, reading a signed action type or evidence class is independent of using the same table for connectivity. Where the usage would coincide, we replace the source, as with genetic support and mechanism direction.

Biomedical databases are individually incomplete, so a missing edge in a verifier's source does not establish that a claim is false. Reconciliation therefore distinguishes active contradiction from absence. An opposite direction, a safety veto, or a negative curated edge may refute a claim; absence produces a deferred verdict and a recorded cross-source gap. Sources that produce frequent gaps are also added to the retrieval layer. The system can then retrieve the union of the available evidence and adjudicate differences, rather than relying on a verifier to arbitrate between two partial databases.

A load-time whitelist excludes datasets that define outcome labels from every verifier. These include curated indication gold standards, held-out repurposing splits, and trial outcome fields. Clinical evidence checks may read registration information available before the evidence cutoff, but never result fields. This restriction prevents the evidence score from recovering the outcome label and preserves the orthogonality of the evaluation axes.

A.4 Evidence score

Let $\mathcal{C}$ be the claims recovered from a trajectory, taken from the structured evidence path when available and from the system's self-cited claims otherwise. Each claim receives an attribution verdict against its cited source, a correctness verdict from the verifier for its claim type, and a credibility level $\ell(c)$ on a five-point ladder. Factuality is the credibility-weighted fraction of claims that survive both checks, and hallucination counts only claims an independent check contradicts:

$$\text{fact} = \frac{1}{|\mathcal{C}|} \sum_{c\,:\,\text{correct}(c)} \frac{\ell(c)}{\ell_{\max}}, \qquad \text{hall} = \frac{|\{c : \text{refuted}(c)\}|}{|\mathcal{C}|}. \tag{1}$$

Support is computed over the load-bearing subset $\mathcal{L} \subseteq \mathcal{C}$ as the product of coverage-weighted quality and a source-convergence factor κ that rewards independent corroboration ($\kappa = 0.75$ for a single

source, 1 for two or more, with a tool confirmation counting as one source):

$$\text{supp} = \left(\frac{1}{|\mathcal{L}|} \sum_{c \in \mathcal{L}} \text{correct}(c) \, \frac{\ell(c)}{\ell_{\max}} \right) \cdot \kappa. \tag{2}$$

The credibility ladder $\ell(\cdot)$ places regulatory labels and randomized trials at the top, followed by observational studies and peer-reviewed reports, with single mined assertions and unresolvable citations at the bottom. A step confirmed by an independent tool is admitted at the observational level or above. Only fabrication, defined here as a claim contradicted by an independent check, counts toward hallucination. A novel hypothesis without direct support is recorded as unsupported and routed for revision. This rule allows repurposing hypotheses to be evaluated without treating a lack of direct evidence as fabrication.

A refuted load-bearing claim leaves $\mathcal{L}$ unsatisfied and lowers support even when every remaining claim is true, because the conclusion still depends on the refuted step. Comparing support with the reported promisingness p gives an over-claim signal, $\text{align} = 1 - |p - \text{supp}|$. The gate blocks a trajectory if it contains a fabricated safety claim, hallucination above a threshold, or evidence dated after the cutoff. It requests revision when a load-bearing step is unsupported or the conclusion is over-claimed, and otherwise proceeds.

A planned refinement extends support from corroboration counting toward a likelihood-ratio form. An edge can be true but uninformative if it would appear for almost any drug, as with highly pleiotropic hubs. The refinement would therefore measure specificity relative to a null. We report the corroboration form because it is the quantity computed by the current implementation.

A.5 Process rubric

The process axis uses 26 items: an integrity gate plus four or five items for each of the six capabilities reported in HDS6. Each item is anchored on a 0–4 scale. Five items are marked critical and cap their capability when violated: state-constraint retention, source integrity, boundary and scope, tool selection, and the integrity gate itself. The gate excludes a trajectory from process scoring if it contains fabricated tool output, an invented identifier, or a command narrated but never issued. A fabricated transcript does not provide a valid basis for scoring the remaining process capabilities.

Self-correction is rated only when the trajectory contains an opportunity to correct an error. Otherwise, the capability remains unrated and is omitted rather than scored as zero. Assigning full repair marks to an error-free run would confound recovery ability with error frequency and could create an apparent capability difference between systems that differ only in how often they err. When verifier feedback is present in a trajectory, the rubric records whether the system acts on it or ignores it. The rubric was calibrated to convergence: calibration terminates when an iteration produces no anchor edits and no criticality changes.

A.6 Effectiveness evaluation

We validate the evidence axis without reference to trial outcomes because it is intended to detect unsound reasoning. We inject errors of known type into trajectories built over the benchmark pairs and measure whether the gate flags them. The language model's verdict is held fixed so that detection can be attributed to the verification layer.

Injected errors are reported in two groups. *Detection* errors comprise fabricated citations and evidence dated after the cutoff. They carry no verifier verdict, so the layer must detect them through its citation resolver and cutoff check. *Propagation* errors comprise flipped mechanism directions, swapped entities, and overclaimed safety statements. They supply the contradiction that an edge verifier would return and test whether a refuted load-bearing step reaches the gate. They do not test the verifiers' ability to find the error. Two control conditions must *not* be flagged: a clean trajectory and a trajectory in which a verifier defers because its source is silent. The second control tests the reconciliation rule of Appendix A.3 end to end.

The 100 public pairs yield 700 trajectories. The layer flagged every injected error in both groups and neither control condition (Table 8). We also disable one check at a time to establish that detection depends on the relevant component. Suppressing the cutoff check drops stale-evidence recall from

1.00 to 0.00, and suppressing the citation resolver drops fabrication recall from 1.00 to 0.00. These mutation tests run in the test suite to ensure that perfect recall remains dependent on the checks being tested.

Table 8: Stratified error injection over the 100 public benchmark pairs, 700 trajectories. Detection errors must be caught unaided; propagation errors test that a verifier contradiction reaches the gate; controls must not be flagged. Intervals are Wilson 95%.

Group	Condition	Flagged / n	Rate (95% CI)
Detection	fabricated citation	100/100	1.000 [0.963, 1.000]
	post-cutoff evidence	100/100	1.000 [0.963, 1.000]
Propagation	flipped mechanism	100/100	1.000 [0.963, 1.000]
	entity swap	100/100	1.000 [0.963, 1.000]
	overclaimed safety	100/100	1.000 [0.963, 1.000]
Control (must not flag)	clean trajectory	0/100	0.000 [0.000, 0.037]
	verifier defers	0/100	0.000 [0.000, 0.037]

We separately test reconciliation by applying the production mapping to every absence label over every pair. None of the 700 cases produces a refutation. An active contradiction, a documented severe overdose, still produces a refutation, confirming that the mapping does not simply defer on all inputs.

Two smaller measurements use live tools rather than scripted verifier verdicts. For each of eight marketed drugs, a scripted model judgment marks a safety claim as confirmed. The live label and pharmacovigilance checks independently overturned that judgment for four of them. The network-proximity check, run live against STRING, scored a matched drug–indication pair at 0.82 ($p = 2.9 \times 10^{-9}$) and a mismatched control at 0.30. It deferred rather than refuting when a drug's targets were not network-proximal to the disease genes.

On 50 real development trajectories, the system's promisingness score was already informative about trial outcomes (AUROC 0.859). Recalibrating it by evidence support left outcome accuracy essentially unchanged (Brier 0.1935 to 0.1954). Support was negatively correlated with prediction error (Spearman -0.206), and the gate flagged 46 of 50 trajectories for revision. Most trajectories cited genuine sources but left at least one load-bearing step corroborated by a single source. These results support the layer's use as a diagnostic for the reasoning failures it targets; they do not establish an improvement in outcome prediction.

A.7 Implementation status

The implemented catalog entries are entity grounding, pathway coherence, genetic support, mechanism direction, toxicity veto, target expression, overdose margin, pharmacovigilance, network proximity, and the aggregate treat-level check. The evidence scorer, reconciliation rule, and offline deterministic tests are also implemented. Binding affinity, signature reversal, gene-set enrichment, and the cell-state fallback are specified with their sources and statistics but are not yet implemented. The target-expression check is validated against recorded Human Protein Atlas responses rather than a live service. The process rubric is implemented, and its results appear in Table 13.

The injection, control, and reconciliation experiments run offline and deterministically as part of the test suite. The suite also includes mutation tests asserting that disabling a check collapses the corresponding recall. In the benchmark mode used for the reported experiments, checks run harness-side and are not exposed to the agent as a callable tool. The tool registry supports that exposure in deployment mode. Reward use of these signals is not part of the reported experiments.

B Development-Set Curation Rules

This section gives more details regarding the rules under which the datasets described in Section 4.1 were built.

B.1 Entity Resolution

Trials are aggregated from ClinicalTrials.gov and the EU Clinical Trials Information System (CTIS) with actual or planned primary-completion dates from 1993 to 2036. We add a trial to our corpus only if its drug and disease resolve to entities in public databases and the pair is tested in an interventional arm, whether for treatment, prevention, or supportive care. A trial registered on both ClinicalTrials.gov and CTIS is counted once: the two records are linked when the ClinicalTrials.gov secondary identifiers list the CTIS trial number, the CTIS record lists the NCT number, or both list the same EudraCT number. This links 2,289 of the corpus's 4,587 CTIS records to ClinicalTrials.gov records, leaving 33,634 unique trials (31,336 ClinicalTrials.gov and 2,298 CTIS-only).

Drugs are normalized against DrugBank (Knox et al., 2024), ChEMBL (Zdrazil et al., 2024), DrugCentral (Avram et al., 2023), ChEBI (Malik et al., 2026), and PubChem (Kim et al., 2025), in that order of preference, and keep one identifier throughout the corpus. A fixed-dose combination is treated as a single entity. Any drug that resolves to a record in one of these databases, whether a small molecule or a biologic, is a repurposing candidate for every indication it is not approved for.

Diseases are normalized to MONDO (Vasilevsky et al., 2022), with EFO (Malone et al., 2010) used only when no MONDO term exactly matches the trial population. The disease assigned to a pair is the condition the trial primarily treats or, for prevention and supportive-care trials, the condition it primarily prevents. Established histologic or biomarker-defined subtypes receive their own disease term, while stage, treatment setting, line of therapy, required comorbidity, and trial-specific biomarker restrictions are retained in the disease string because they define distinct indications; severity and age group alone do not. A qualified subpopulation becomes a separate pair once a posted result or approval is specific to that population; until then, unresolved records for the same drug and disease term are grouped into one pair. A record is assigned to a pair only if its population matches the pair definition, and trials with cohorts that have separate primary analyses define a separate pair for each cohort.

Attribution. Using an LLM judge, each record is attributed to the drug(s) the trial is designed to test, judged from its arms, primary objectives, and registry description rather than from registry role codes. A regimen held constant across arms while the trial contrasts another agent is background, whether or not it is approved, and is not attributed; the added agent is the tested drug, and a standard-of-care comparator is a control. A constant regimen is attributed only when the trial tests the regimen itself (for example, its dose, schedule, or use in a new population). In a head-to-head trial each non-background active agent defines its own pair, and a single-arm trial tests every non-standard agent it administers. A record joins a pair only when it tests that pair's exact drug entity.

Approvals are matched to the pair's exact drug entity and indication (disease term plus qualifier), using the earliest indication-level approval by a major regulator (FDA, EMA, PMDA, NMPA, CDSCO). Approval for a narrower indication, or of a single component of a tested combination, is not considered an approval for that pair.

B.2 Outcome Grading

An LLM judge (Claude Opus 5) extracts a candidate outcome label from each trial's posted results, and a human annotator reviews and corrects every extraction. We grade the prespecified primary efficacy, safety, and validated-biomarker objectives that apply to the pair's population. An objective is met when the registry's posted analysis of the contrast that isolates the target drug (against the same background without it, or against the registered comparator) meets its prespecified criterion under the trial's own multiplicity and non-inferiority conventions. When the registry posts no such conventions, we test the isolating contrast on the posted summary statistics (Welch's t-test, Fisher's exact test, or a z-test on least-squares means) with a two-sided test at $\alpha = 0.05$, requiring the effect to be in the prespecified direction. Such computed tests decide 6 of the 45 decided trials in the public set. In a factorial trial, the target-drug effect is graded from the factorial analysis or a pooled comparison across the levels of the other factor. A safety objective is graded only against a registered criterion or comparative hypothesis, feasibility objectives are never graded, a dose-ranging objective is met if it is met at any dose, and a single-arm objective without a registered threshold is graded by the reviewed judge assessment of whether the observed activity is meaningful. A Phase II-or-later record whose only primary objectives are non-gradable is labeled *Ongoing* and still counts as evidence of clinical advancement.

Table 9: Stated training-data cutoffs of the evaluated backbones relative to the benchmark cutoff of January 1, 2026, as published in each developer's model documentation. "Not stated" marks a developer that publishes no cutoff; every such backbone was released after the benchmark cutoff. †Reported in third-party launch coverage, not in the developer's documentation.

Backbone	Stated training cutoff	Released	Before benchmark cutoff
Claude Opus 5	May 2026	Jul 24, 2026	No
GPT-5.6-sol	Feb 16, 2026	Jul 9, 2026	No
GPT-5.5	Dec 1, 2025	Apr 23, 2026	Yes
Kimi K3	Not stated	Jul 16, 2026	Unknown
Qwen3.8 Max	Not stated	Aug 3, 2026	Unknown
DeepSeek V4 Pro	Not stated	Apr 24, 2026	Unknown
GLM-5.2	Mar 2026†	Jun 2026†	No

A trial is labeled *Success* if all applicable primary objectives are met, *Failure* if none are met or it stops early for any reason, *Mixed* if they split, and *Ongoing* if no applicable primary result is extractable.

The registry's phase designation is taken as given, with two clarifications. First, a combined Phase I/II or II/III trial counts at its later phase only if the later portion enrolled, and at its earlier phase otherwise. Second, a Phase IV designation denotes a postmarketing study of a drug already approved for some indication (Code of Federal Regulations). Every Phase IV record in the public set tests a marketed drug in an indication for which it is not approved; because the drug has already completed Phases I–III and regulatory review in another indication, we place such records on the tier above Phase III (Table 2). Unresolved Phase I trials are not used as evidence: they neither date nor grade a pair, nor establish advancement.

The records map to targets as in Table 2, and an approval on or after the cutoff enters the pair as a record at 1.00. A pair takes its most favorable record: the highest floor, preferring at the same floor a lower bound over an exact target and a decided success over an unresolved trial, so a post-cutoff approval supersedes every trial. The grade values follow the empirical distribution of 10,427 programs completed in the training corpus (Section 2).

B.3 Data Leakage

Evidence dates and cutoff. Each trial is dated solely by its registry record: a concluded trial by the date the registry first posted results, a trial decided by its termination reason by the update that posted the termination, and an unresolved Phase II-or-later trial by the latest posting that establishes its status. Registry records reflect a snapshot taken on September 20, 2026. A pair is eligible if its most recent record, ignoring unresolved Phase I records, is dated on or after January 1, 2026, and is excluded if its drug was approved for the indication before that date. Earlier records of an eligible pair are kept as history and can still set its grade: dates decide eligibility only, never the grade.

Model training data. Choosing the cutoff trades corpus size against memorization: an earlier date leaves fewer eligible pairs, while a later one lets more outcomes into model training data. We set it at January 1, 2026, just after GPT-5.5's stated training cutoff of December 2025. The other backbones may have seen some early-2026 registry postings: Claude Opus 5, GPT-5.6-sol, and GLM-5.2 state cutoffs between February and May 2026, and the rest state none but were released after the cutoff (Table 9). GPT-5.5 therefore serves as our reference backbone with no stated exposure.

Earlier public disclosure. A trial result can appear in a press release, abstract, or journal article before the registry posts it. Such publications never date or grade a trial. Of the 100 public pairs, 8 have a registry-linked publication dated before the cutoff.

Tool access. In the with-environment condition, every retrieval source (knowledge graphs, literature index, and tool databases) is a frozen snapshot dated before the cutoff, and no live web access is available.

Table 10: Component matrix and performance of the ten completed combinatorial ablation recipes. √/× denote an available/unavailable component. Raw scores are seed means; normalized scores are reported as mean ± sample SD across three complete 100-example seeds. Trials and Tables refer to an date gated Janurary 2026 AACT snapshot used only in this sensitivity design

Recipe ID	Wiki	Trial	Tables	Bio	KG	PF	Seeds	Raw score	Normalized
0	×	×	×	×	×	×	3	95.86	68.34 ± 5.12
1	√	√	√	√	×	×	3	95.59	66.30 ± 3.67
2	√	√	√	×	×	×	3	95.75	67.47 ± 7.46
3	√	√	√	√	√	×	3	96.48	73.08 ± 5.82
4	√	√	√	√	×	√	3	96.73	74.97 ± 1.16
5	×	√	√	√	√	√	3	94.55	58.34 ± 4.17
6	×	√	√	√	×	×	3	95.44	65.10 ± 4.33
7	×	×	×	√	√	√	3	94.03	54.34 ± 3.07
8	×	√	√	×	√	√	3	95.37	64.56 ± 9.25
9	√	√	√	√	√	√	3	94.99	61.66 ± 3.32

C Ablation Study with Bayesian Inference

Table 10 gives the complete component matrix and observed performance for the ten completed three-seed recipes used in the Bayesian ablation study. (Section 4.2).

For each reported value i, let y_i be normalized score, $m[i]$ its backbone, x_{ik} the presence indicator for component k, g_i the Full-v2 protocol indicator, and $e_i = \mathbb{1}$[at least one environment component is enabled]. GPT-5.6-sol is the reference backbone ($a_{\mathrm{GPT}} = u_{\mathrm{GPT}} = 0$). We fit

$$y_i \mid \mu_i, v_i \sim \mathcal{N}(\mu_i, v_i), \tag{3}$$

$$\mu_i = \alpha + \sum_{k \in \mathcal{K}} \beta_k x_{ik} + \delta g_i + \gamma_{s_i} + a_{m[i]} + u_{m[i]} e_i, \qquad \gamma_0 = 0. \tag{4}$$

The a_m terms absorb backbone baselines and u_m terms backbone-specific environment-response deviations. They are nuisance controls, not reported scientific effects. To avoid differential fixed-scale regularization, every regression coefficient in the equation uses the same weakly informative prior, $\mathcal{N}(0, 20^2)$. For recipe-seed observations, v_i is the squared across-seed standard deviation of the relevant recipe (lower-bounded at one score point); for main-table summaries, it is the reported three-seed standard error squared. Thus main-table rows enter as summary observations rather than pseudo-seeds. The seven matched stable main-table backbone pairs add fourteen summary observations. Their environment condition is coded as Literature Wiki, BioNeMo, biomedical KG, and Meta-Pathfinde The combinatorial recipe matrix supplies the contrasts that separate the components.

D Construction of the Literature Wiki/KG Provider

The Literature Wiki/KG is designed to synthesize a corpus of approximately 785K drug-repurposing papers while preserving traceability to the source evidence. A conventional LLM-maintained Wiki updates entity pages sequentially, which creates a serialization bottleneck and makes earlier page content sensitive to document order. We instead use the following parallel, evidence-first construction.

Map: paper-level claim extraction. Each paper is processed independently by a versioned extraction prompt. The model emits structured claims identifying the participating drug, disease, target, mechanism, and other biomedical entities, together with the claim's direction and its source reference. Long inputs are handled with a bounded whole-document or head–tail view. Because the map stage performs one independent call per paper, the corpus can be sharded across workers without shared mutable pages.

Shuffle: canonicalization and evidence buckets. The extracted claims are normalized to stable, release-local keys and shuffled into two kinds of buckets: relation buckets for drug–disease hypotheses and entity buckets for drugs, diseases, targets, and mechanisms. Alias mappings are recorded alongside the registry. This stage is deterministic and runs offline, so regrouping or updating claims does not require model inference. The current release uses lexical canonicalization, with optional clustering of near-duplicate aliases; full ontology grounding is not assumed.

Reduce: relations before entities. Relation buckets are reduced first. The model synthesizes all claims about a drug–disease pair into a hypothesis page that preserves supporting and contradictory evidence and its references. Entity buckets are built only after these relation summaries are available; an entity page therefore combines its incident relation summaries with claims that are not specific to one drug–disease pair. This ordering avoids repeatedly injecting every raw claim into high-degree entity pages and produces consistent summaries between the relation and entity views.

Both reductions use a hierarchical combiner tree. When a bucket contains more than a configured number of evidence rows, it is partitioned into slices; each slice is summarized into a partial, and the partials are recursively combined into the final page. Each bucket assigns global reference identifiers before it is split, and every intermediate and final output is checked to use only references present in its input. The procedure therefore avoids both silent evidence truncation and unconstrained citations for hub entities such as frequently studied drugs or genes.

Assembly and serving. The reducer emits a versioned JSONL substrate containing page metadata, alias mappings, structured drug–disease edges, and contradiction records. This substrate is the authoritative representation. Markdown entity and relationship pages, indexes, backlinks, and update logs are rendered deterministically from it rather than edited by hand. The serving provider loads `pages_meta.jsonl`, `alias_map.jsonl`, and `edges.jsonl` to resolve titles and aliases and to expose directed `may_treat` adjacency. It additionally derives Wiki-link adjacency from links in rendered pages. Rich page text and PubMed citations remain off the search frontier and are returned through the enrichment operation only for selected node and edge identifiers.

Incremental updates and limitations. When new papers arrive, completed map shards are reused, the inexpensive shuffle is rebuilt, and only relation or entity buckets whose contents changed are re-synthesized. Assembly then re-renders the affected pages and records create, update, or retire events. The present pipeline should not be described as fully ontology-grounded: grounding Wiki entities to MeSH, MONDO, ChEMBL, and HGNC identifiers, LLM-based adjudication of contradiction records, paper-stub pages, and a dedicated mid-training rendering pass are documented subsequent stages rather than assumptions of the current construction. Meta-Pathfinder therefore treats Wiki identifiers as provider-local and performs conservative cross-provider reconciliation at query time.

E Reformulation Procedure Design

E.0.1 Definition

Drug reformulation goes beyond asking whether a drug may work: it specifies how the candidate should be developed, for whom, and under which formulation, delivery, biomarker, and combination strategy. These outputs are prospective design hypotheses and must remain tied to the available evidence.

Unlike the repurposing score, which summarizes the expected clinical-development position of a drug–disease pair, reformulation does not reduce to a single stage label. Its purpose is to specify actionable design choices and the evidence and open questions needed to evaluate them.

The reformulation subtask has four linked design dimensions, summarized in Table 11:

Table 11: Four linked design dimensions of the drug-reformulation subtask.

Design dimension	Required output
Biomarkers	Identify biomarkers for efficacy prediction, patient selection, and treatment-response monitoring, with mechanistic rationale and supporting clinical evidence.
Patient stratification	Identify the disease subtypes, molecular characteristics, genetic backgrounds, and clinical features that define the patients most likely to benefit.
Formulation and delivery	Recommend dosage form, route of administration, targeted delivery, or sustained-release strategies that could improve bioavailability, safety, and clinical feasibility.
Combination therapy	Recommend combinations with plausible synergy by integrating mechanisms of action, signaling pathways, and drug–drug interactions, while making expected benefits and risks explicit.

E.0.2 METHOD

We formulate reformulation procedure design as constrained structured generation conditioned on a drug–disease hypothesis and the evidence accumulated during its investigation. For each candidate, the agent assembles an evidence state containing the drug's known pharmacology and safety profile, the proposed disease mechanism, relevant clinical-trial protocols and eligibility criteria, and any mechanism paths recovered from the knowledge network. It then generates four linked fields: target population, biomarker strategy, combination therapy, and formulation or delivery. These fields are not optimized independently: the population must be consistent with the disease stage and available eligibility evidence; a biomarker must be assigned a concrete predictive, stratification, or response-monitoring role; a proposed combination must specify both mechanistic complementarity and overlapping risks; and a formulation change must address an identified limitation in exposure, tissue access, safety, or administration. The agent first drafts the complete procedure from the shared evidence state and then revises fields for which structured clinical context is available; in the current implementation, retrieved eligibility records are used to refine the target-population specification. Each recommendation separates source-supported facts from prospective design hypotheses and records the missing evidence or validation step needed to test it. The resulting output is therefore an auditable development plan for the proposed indication, rather than a claim that the reformulation itself has established efficacy.

E.0.3 EVALUATION

We separately evaluate whether a system can propose a clinically actionable development procedure for a drug–disease pair. The system predicts four structured fields: target population, biomarker strategy, combination therapy, and formulation or delivery. Reference values are derived from registered trial protocols. In the language-model-only condition, these fields and their source records are withheld during prediction. An automated semantic evaluator compares each predicted field with its reference on a 1–5 scale, where 1 denotes an irrelevant or contradictory proposal, 3 denotes a directionally correct but incomplete proposal, and 5 denotes agreement on the key clinical content. Field means are computed over examples with an available reference value, and the macro-average gives equal weight to the four field means.

For each item with a non-null reference, deterministic overlap and route checks are reported separately as auxiliary diagnostics.

E.0.4 RESULTS

Table 12 compares a language-model-only condition (LM only) with a knowledge-augmented condition (LM + context). The augmented condition revises the initial model prediction using structured eligibility context retrieved from registered clinical-trial records. This context is available at sufficient coverage only for population specification; the biomarker, combination, and formulation/delivery predictions are therefore identical across conditions. Because the retrieved eligibility record is also the source from which the population reference is derived, the population comparison measures evidence retrieval and synthesis rather than fully prospective procedure generation. Population improves by 0.89 points, increasing the macro-average from 3.72 to 3.94. Together, the

Table 12: Drug repurposing and reformulation procedure-design results. Scores are mean semantic-evaluator ratings on a 1–5 scale. LM + context differs from LM only for population, the field for which structured trial context is available.

Field	LM only	LM + context	Δ
Population	2.89	3.78	+0.89
Biomarker	4.00	4.00	0.00
Drug combination	3.75	3.75	0.00
Formulation/delivery	4.22	4.22	0.00
Macro-average	3.72	3.94	+0.22

Table 13: HDS6 capability scores on matched drug–disease cases. Condition rows report mean ± population standard deviation on the 0–4 scale; NR and missing capability ratings are omitted. Absolute gain is the mean paired within-case difference (DORA minus LLM-only) among cases rated in both conditions. Repair gains are unavailable because repair was not observable in the LLM-only trajectories.

Backbone	Condition	C	E	A	S	T	R
GPT-5.5	LLM-only	1.62 ± 0.21	1.52 ± 0.32	2.35 ± 0.65	2.72 ± 0.88	1.00 ± 0.00	—
	DORA	2.76 ± 0.72	3.10 ± 0.62	2.69 ± 0.64	2.96 ± 0.65	3.64 ± 0.69	3.45 ± 0.79
	Absolute gain	$+1.13$	$+1.58$	$+0.34$	$+0.24$	$+2.64$	—
GPT-5.6-sol	LLM-only	1.67 ± 0.18	1.42 ± 0.32	2.48 ± 0.72	2.85 ± 0.79	0.99 ± 0.15	—
	DORA	2.81 ± 0.82	3.10 ± 0.65	2.72 ± 0.62	3.09 ± 0.62	3.56 ± 0.79	3.83 ± 0.35
	Absolute gain	$+1.14$	$+1.68$	$+0.24$	$+0.24$	$+2.57$	—

outcome, trajectory, and procedure-design evaluations assess three complementary capabilities: selecting promising therapeutic candidates, producing evidence-grounded reasoning, and specifying clinically actionable development strategies.

E.0.5 HDS6 Trajectory Analysis

We use HDS6 to characterize the quality of the scientific reasoning process in addition to final-answer accuracy. HDS6 (Wang et al., 2026) rates six dimensions on a 0–4 scale: long-horizon state coherence (C), evidence fidelity (E), hypothesis management (A), boundary and failure reasoning (S), tool use and execution-state management (T), and self-correction under verification (R). Scores of 0, 1, 2, 3, and 4 correspond to absent, flawed, insufficient, good, and excellent behavior, respectively. An automated process evaluator scores the recorded reasoning and tool interactions with the held-out outcome hidden. Capability scores retain the rubric's critical-failure caps. A separate integrity gate excludes trajectories that are too incomplete or malformed for reliable process scoring, and a capability marked not rated (NR) is omitted rather than assigned a score of zero. Table 13 reports capability means, dispersion, and paired condition gains.

Across backbones, the environment most strongly improves tool use, evidence fidelity, and long-horizon coherence, while gains in hypothesis management and scope control are smaller. This pattern suggests that the environment primarily helps systems ground claims, preserve state, and execute evidence-seeking workflows. The low no-env tool scores reflect the deliberate absence of executable tools, not failed attempts to use them.

GPT-5.5 and GPT-5.6-sol exhibit similar with-env profiles, indicating that the process-level benefits are stable across the two GPT backbones. Repair scores are conditional on an observable correction opportunity and are unavailable for the no-env trajectories, so they should not be treated as directly comparable coverage-matched estimates. The HDS6 findings are descriptive and remain subject to the survivor-selection limitation above.