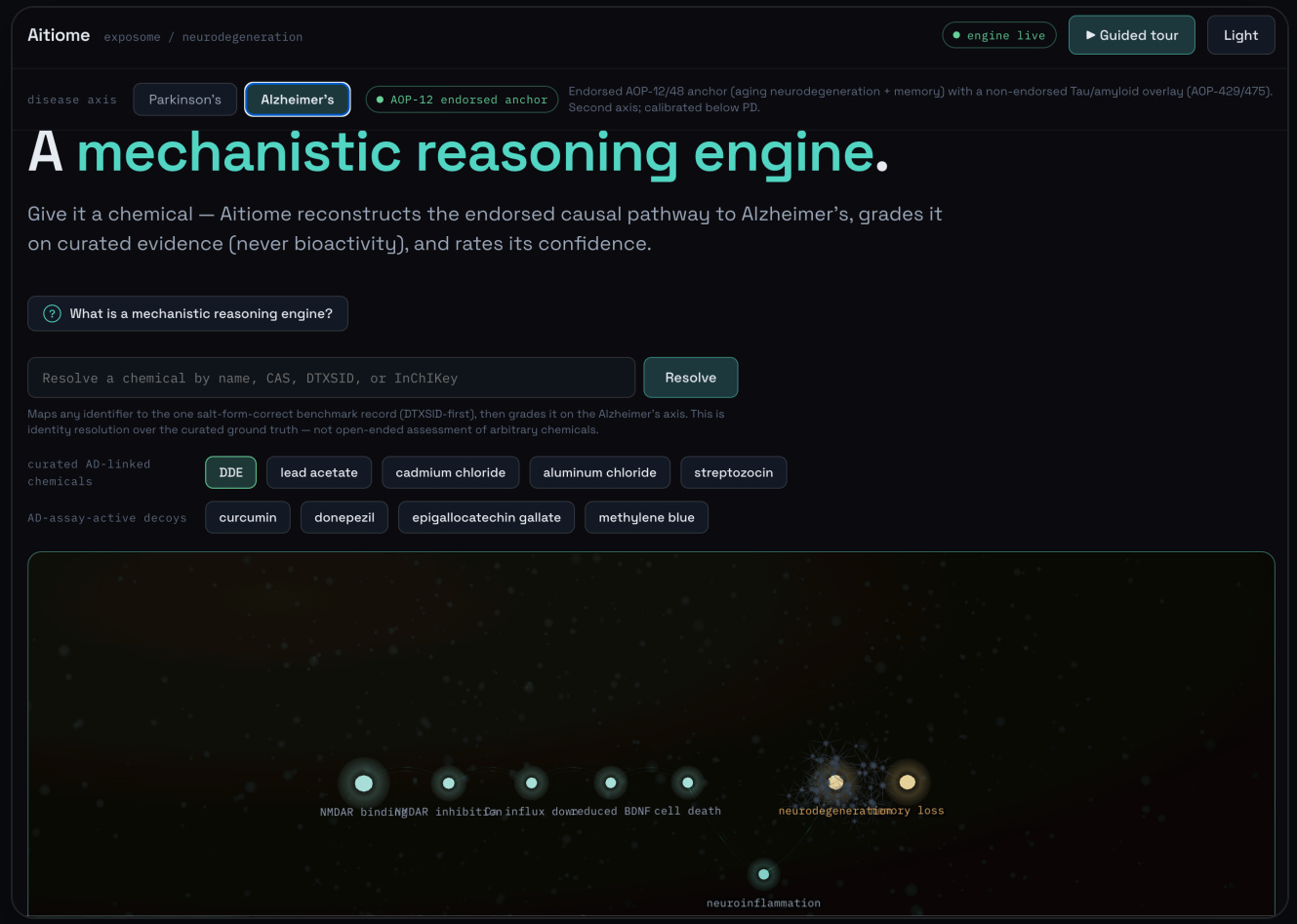
Aitiome

A mechanistic reasoning engine for the environmental exposome of neurodegeneration.

It recovers the real ones, resolves to the real neurons, and isn't fooled by the imposters.

Submitted by **Jon Radoff** for the **Built with Claude: Life Sciences** Hackathon.

aitiome.fly.dev / Built with Claude: Life Sciences



WHAT IT IS

Give it a chemical; it reconstructs the endorsed causal pathway to Parkinson's or Alzheimer's, graded on **curated evidence - never bioactivity**.

WHO IT'S FOR

Exposome researchers and mechanism-focused labs - the community that called for AI linking chemicals to disease *mechanism*.

WHY IT'S HARD

Thousands of chemicals are bioactive; almost none are curated neurotoxicants - and **bioactivity scores worse than chance** at telling them apart.

HOW IT'S MADE

Deterministic Go engine; **Claude Opus 4.8** narrates the evidence, **built with Claude Code**, with an **MCP server** for agents.

Try it live at **aitiome.fly.dev**

WHAT WE ACCOMPLISHED, UP FRONT

A validated engine - and a methods result

13/13 PD recovered (12/12 AD) **6/6** decoys rejected **0** false pos + neg **live** + MCP

Why what we built matters

The field asked for AI that links chemicals to *mechanism*, not another activity screen. Aitiome is that validated prototype: it recovers the known neurotoxicants on the endorsed pathway, **rejects the bioactive decoys that fool activity models**, calibrates its confidence, and maps the discovery limits honestly instead of overclaiming.

The methods result

Studying how Claude should synthesize evidence, **adversarial RLM surfaced ~10x more counter-evidence than RAG and ~2.2x more than the strong RAG+ baseline** - matching RAG+ on yield while running degraded. Decomposition makes Claude a measurably better skeptic: the property a calibrated-honesty tool needs most.

THE PROBLEM

What causes sporadic Parkinson's and Alzheimer's?

The large majority of Parkinson's and Alzheimer's is **sporadic** - not inherited. A major, understudied contributor for the rest is the **environmental exposome** - the pesticides, metals, and industrial chemicals we meet over a lifetime, interacting with genetics, aging, and vascular biology. In 2024 the field's leaders (Miller, Barouki, Samieri; *Nature Neuroscience*) called for AI that connects chemicals to disease *mechanism*.

Which chemicals have mechanistically supported links to neurodegeneration, and through what biological pathway? That is the open question this project takes on.

WHY IT'S HARD

Bioactive is not the same as neurotoxic

Tens of thousands of chemicals are **bioactive** - they light up assays. Very few have curated evidence for human neurodegeneration. The entire difficulty is telling a true neurotoxicant apart from a compound that is merely bioactive.

This is the exact failure mode of activity-based screening: it flags everything that *does something*. Beat that, and you have something real. Fail it, and you have a watchlist no scientist can trust.

The whole project is built to win that one discrimination.

WHAT WE SET OUT TO BUILD

A discovery engine for novel drivers

The original pitch: an AI that scans the environmental chemical universe and surfaces *novel* candidate drivers of neurodegeneration, validated by recovering the known neurotoxicants.

Before writing code, we did a multi-round data reconnaissance to check whether that was actually possible on public data. It changed the project - and that is the interesting part.

DISCOVERY, SCOPED INTO ENGINEERING + SCIENCE

What the data supports - and what it doesn't yet

THE GAP WE FOUND

A full untargeted-discovery engine needs data public sources can't yet supply for this chemical class - seven independent axes tested, all coverage- or confounder-killed. The discriminating signal is **curated mechanism**, not assay activity (bioactivity scores **worse than chance** against the decoys).

SO WE DECOMPOSED IT

We **built the components the evidence supports** - a validation engine and a candidate triage pipeline - and **specced the wet-lab experiments** needed to complete the discovery half. Engineering where the data is ready; a mapped science agenda where it isn't.

The same discovery goal, scoped to what the evidence can honestly support.

WHAT AITIOME IS

Three co-equal goals

01

Validation that works

Recover the known neurotoxicants on the endorsed pathway; reject the adversarial decoys.

02

A signature visualization

The recovery-and-specificity reveal resolving into the vulnerable dopaminergic neurons.

03

Calibrated framing

Confidence tiers on every result; the discovery limits shown, not hidden.

THE CORE DISCIPLINE

The recovery rule

```
positive  ⇔  ( CTD curated Parkinson's DirectEvidence )  OR  ( registered neuro-AOP stressor )
```

Curated signals are diagnostic.
Two independent curation efforts. A fixed logical rule, no fitted parameters to overfit.

Assay activity is corroboration only.
Anti-diagnostic here - the engine **never** gates a positive call on bioactivity.

THE SIGNATURE VISUAL

The endorsed cascade, grounded, resolving into the vulnerable neurons

MIE complex-I inhibition → mitochondrial dysfunction → nigrostriatal dopaminergic degeneration → parkinsonian deficits. Grounded in MitoCarta Complex-I subunits and the Kamath SOX6/AGTR1 vulnerable dopaminergic neurons.

THE RESULT, ON THE RECONNAISSANCE GROUND TRUTH

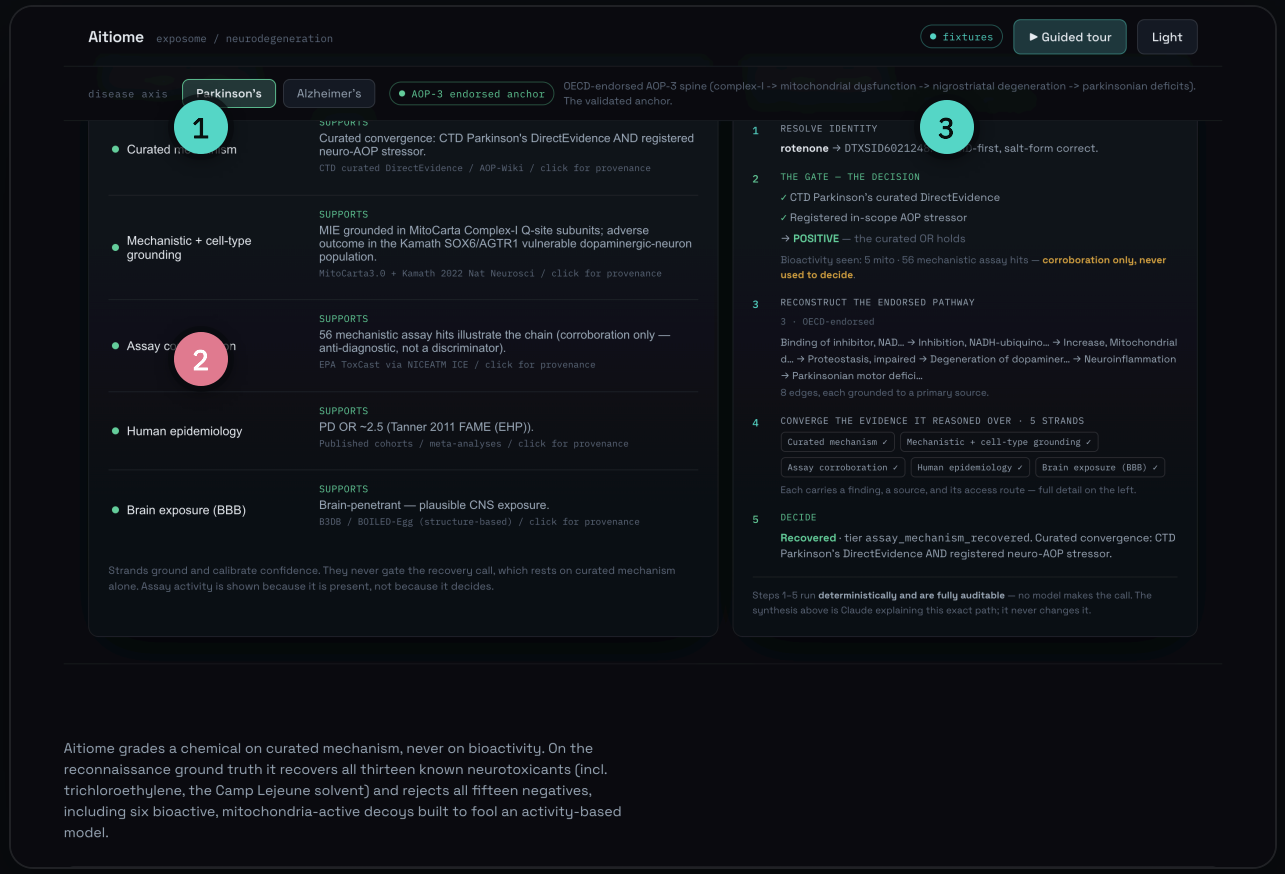
Zero errors

13/13	15/15	6/6	0
known neurotoxicants recovered	negatives correctly rejected	bioactive decoys not fooled	false positives + false negatives

13 Parkinson's positives (6 assay-recovered + 7 curated-anchored, incl. trichloroethylene / Camp Lejeune) - 15 negatives, including 6 mitochondria-active adversarial decoys built to fool an activity model.

A curated adversarial benchmark, not a population-generalization study. Recovery is the sanity check; the **6/6 decoy rejection** is the contribution - always read the two together.

Convergent evidence, and the pathway reconstructed

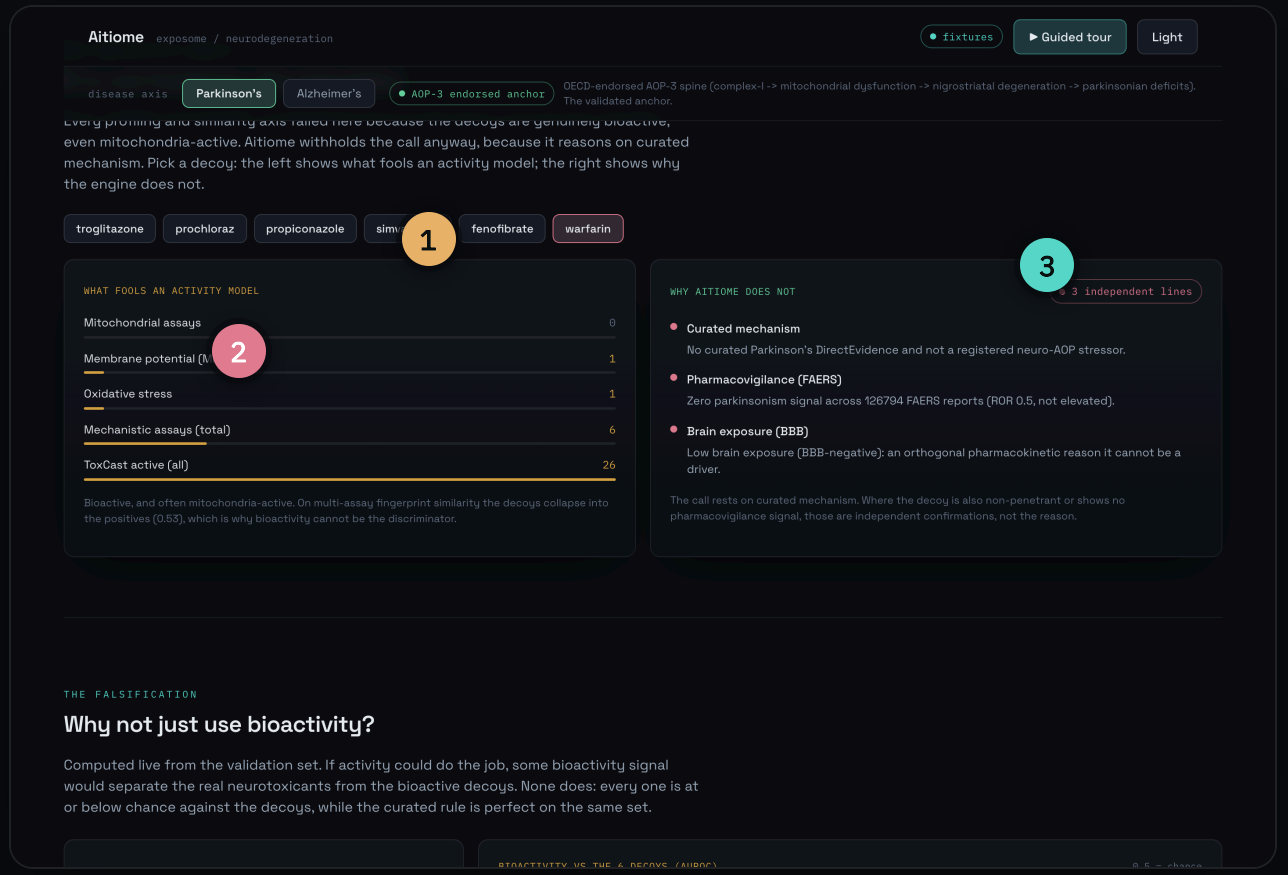


Aitiome grades a chemical on curated mechanism, never on bioactivity. On the reconnaissance ground truth it recovers all thirteen known neurotoxicants (incl. trichloroethylene, the Camp Lejeune solvent) and rejects all fifteen negatives, including six bioactive, mitochondria-active decoys built to fool an activity-based model.

1. Five convergent evidence strands, each with provenance
2. Assay activity is corroboration only - never a discriminator
3. The endorsed AOP, reconstructed edge by edge

THE SPECIFICITY CENTERPIECE, LIVE

It is not fooled by the imposters



1 Pick any bioactive decoy

2 What lights up an activity-based model

3 Rejected on three independent lines at once

THE FALSIFICATION - COMPUTED LIVE FROM OUR OWN DATA

Why not just use bioactivity?

If activity could do the job, some signal would separate the 13 positives from the 6 decoys. None does - every one is at or below chance (0.5).

Curated rule on the same set: **perfect (0 errors)**. The decoys are, if anything, more bioactive than the real neurotoxicants.

Prior chemical-disease ML reports 90%+ AUROC - but on drug-keyed, well-covered space, never adversarial environmental decoys. ToxCast covers only ~40% neural-relevant targets (Mack 2024): the structural reason activity fails here.

ANSWERING THE SHARPEST CRITIQUE, EMPIRICALLY

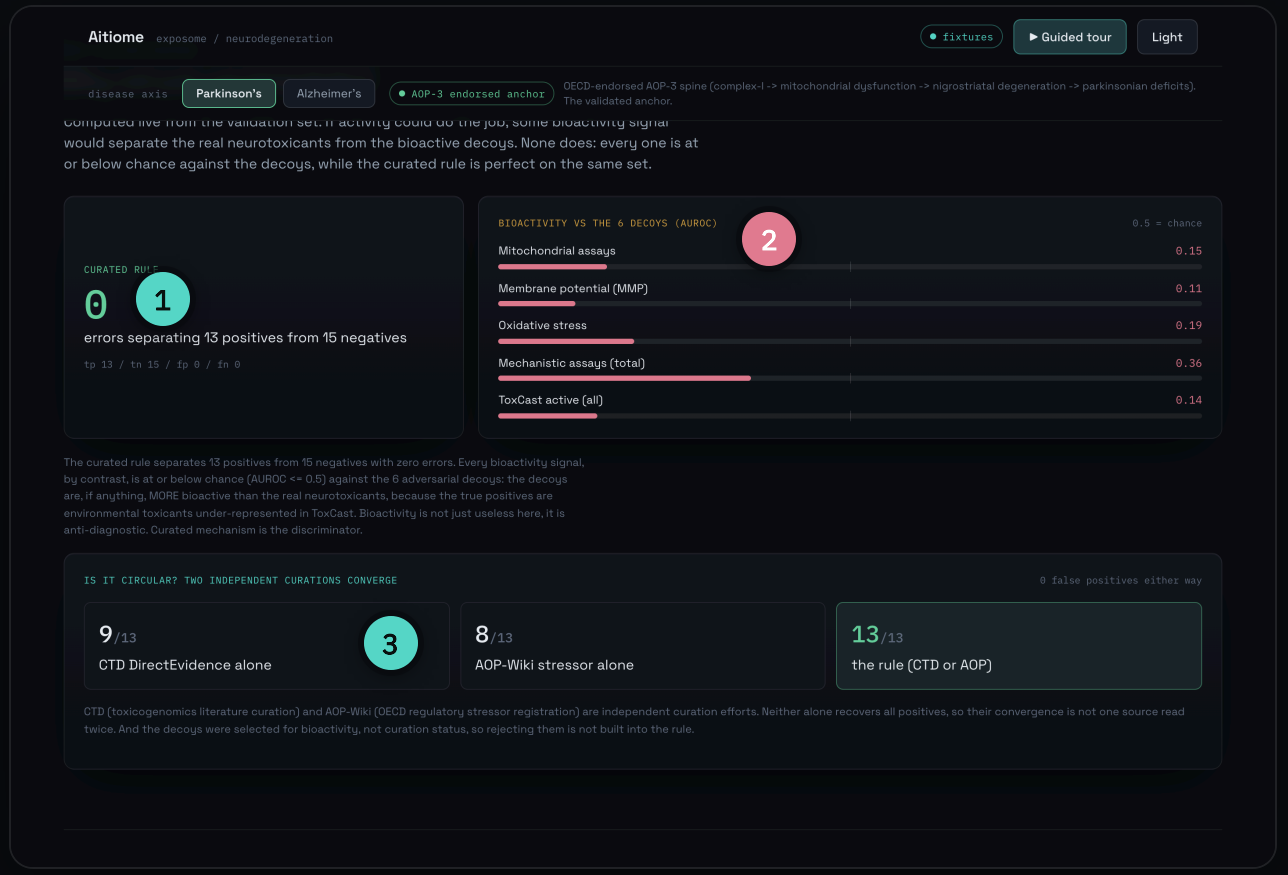
Is it circular? Two independent curations converge

9/13 CTD DirectEvidence alone + 8/13 AOP-Wiki stressor alone → 13/13 the rule (CTD or AOP)

CTD (toxicogenomics literature) and AOP-Wiki (OECD regulatory) are independent efforts; neither alone suffices, so this is not one source read twice. Both are **0/15** false-positives. And the decoys were selected for bioactivity, not curation status - so rejecting them is not baked in.

NOT A SLIDE - COMPUTED LIVE FROM OUR OWN DATA

The whole argument, running in the product



1 Curated rule: zero errors on the set

2 Every bioactivity signal at or below chance

3 Circularity answered: two independent curations converge

THE LIMITS, MAPPED

Where AI-driven discovery works, and where it does not

LINCS L1000 COVERAGE-KILLED	EPA HTTr COVERAGE-KILLED	EPA HTPP COVERAGE-KILLED	Structure / QSAR CONFOUNDER-KILLED
ToxCast fingerprint CONFOUNDER-KILLED	ComptoxAI / AIzKB COVERAGE-KILLED	Neural-specific subset QUALIFIED LEAD ●	Boltz-2 Q-site CONDITIONAL LEAD ●

Seven axes coverage- or confounder-killed. Two unproven leads remain (reported with explicit N). Discovery is a map, never a predictor.

THE HONEST FORM OF THE ORIGINAL GOAL

Discovery, done honestly: a triage queue

VALIDATOR → PIPELINE

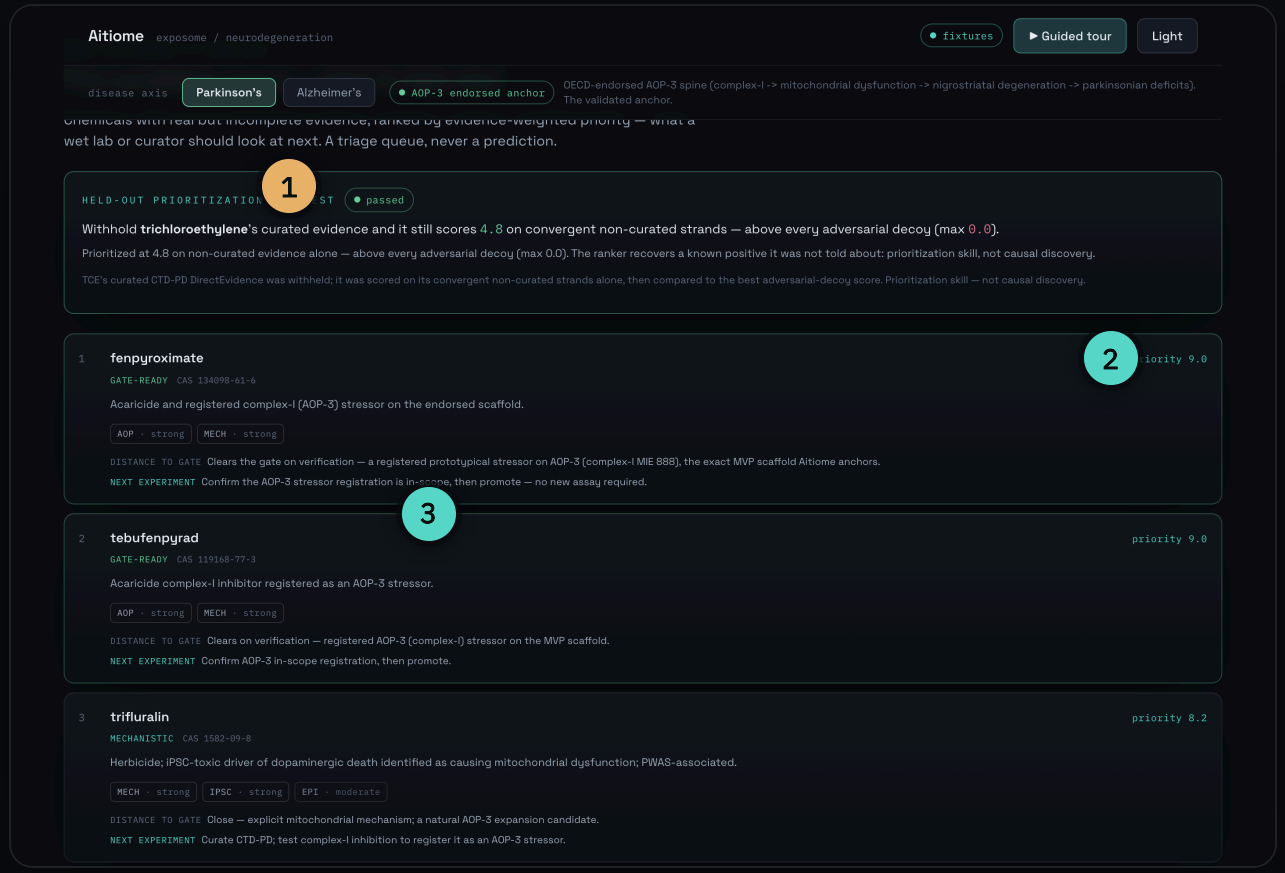
The original pitch was discovery. Its honest form is a **triage queue**, not a predictor: chemicals with real but incomplete evidence, ranked by an **evidence-weighted priority** built only on non-bioactivity strands. Each carries a distance-to-gate and a **recommended next experiment** for a wet lab.

THE DISCIPLINE, KEPT

The ranking never decides - **only the curated gate promotes**. The six adversarial decoys are carried as a permanent control and **rank last (0)**. A **held-out backtest** recovers a known positive (TCE / DDE) from non-curated evidence alone, above every decoy: prioritization skill, not causal discovery.

THE CANDIDATE PIPELINE, LIVE

A ranked queue that tells a lab what to test next



1. Held-out backtest: a known positive recovered from non-curated evidence, above every decoy
2. Evidence-weighted priority - gate-ready AOP stressors rank top
3. Each candidate: distance-to-gate + the recommended next experiment

A SECOND AXIS, THE SAME DISCIPLINE

Alzheimer's — calibrated below Parkinson's

THE AD ARM

The identical curated-diagnostic engine, anchored on the **endorsed AOP-12/48** (aging neurodegeneration + memory; registered stressor: lead) with a non-endorsed Tau/amyloid overlay. Recovers curated AD-linked chemicals (DDE, cadmium, lead), rejects the imposters, and resolves into disease-associated microglia (TREM2/APOE).

THE DECOYS ARE DRUGS AND SUPPLEMENTS

The compounds most active on AD assays are symptomatic **drugs** (donepezil, galantamine - cholinergic; methylene blue - investigational anti-tau) and dietary **polyphenols** (curcumin, EGCG). An activity model would flag them. Aitiome does not - no curated causation.

Same rigor; confidence calibrated to the evidence.

THE TWO AXES, COMPARED

The same method, calibrated to the evidence.

dimension	Parkinson's	Alzheimer's
Scaffold (OECD)	●●● AOP-3 endorsed; EFSA-expanding complex I-IV family	●●○ AOP-12/48 endorsed + non-endorsed overlay
Curated recovery	●●● 13/13, 0 err	●●● 12/12, 0 err
Predictive power / falsification	●●● quantified: activity at or below chance	●○○ qualitative; assay-AUROC pending data
Circularity defense	●●● two curations converge (9/13 + 8/13)	●○○ leans on CTD alone (~11/12)
[illegible]	●●● [illegible]	●●○ [illegible]

Shared: KE-188 neuroinflammation bridges both AOPs; lead is positive for both; mitochondrial dysfunction is an unlinked gap, shown not hidden.

WHERE WE SIT, MEASURED

Prioritization is prior art. The AI approach is ours.

PROTON's AI vs ours

PROTON (2025) is a learned model whose prediction is the output. Aitiome inverts the role of AI: Claude reconstructs the pathway and grades the evidence, but a **curated gate, not the model, makes the call.**

What the data shows

PROTON shares our no-bioactivity discipline and **converges** - endosulfan, dicofol and naled (its held-out top pesticides) are all in our PD queue. ENRICH uses bioactivity and **diverges** - 0 of our 43 reach its top-250, and it ranks our decoy propiconazole #6.

PROTON and Aitiome are complementary - it generates discovery hypotheses, we govern and audit the evidence behind them. The queue itself is prior art (ToxPi, IATA, ENRICH, PROTON); the bundle is ours: bioactivity excluded on evidentiary grounds, a transparent index, ranking separate from a curated gate, decoys as a live control.

GROUNDED IN ESTABLISHED MECHANISM

Convergent with the mechanistic literature, node by node

Parkinson's spine

The mitochondrial-dysfunction → dopaminergic-degeneration edge we reconstruct is the **Nakamura** lab's own work (Berthet 2014; CHCHD2, Sci Adv 2025); the parkinsonian-deficit outcome is the **Kreitzer** lab's basal-ganglia circuit (Kravitz 2010, Nature).

Alzheimer's arm

Our neuroinflammation + BBB node is **Akassoglou's** fibrin→microglia mechanism (Merlini 2019; Mendiola 2023); the APOE→microglia node is **Huang's** APOE4 work (2024-25); TREM2→network is **Mucke** (Das 2021).

The field owns the endogenous mechanism; Aitiome adds the exposome layer.

We also pre-empt Nakamura 2008 (complex-I not strictly required) - exactly why we never gate on activity and model paraquat via redox-cycling AOP-593.

THE EXCLUSIONS ARE THE DISCIPLINE

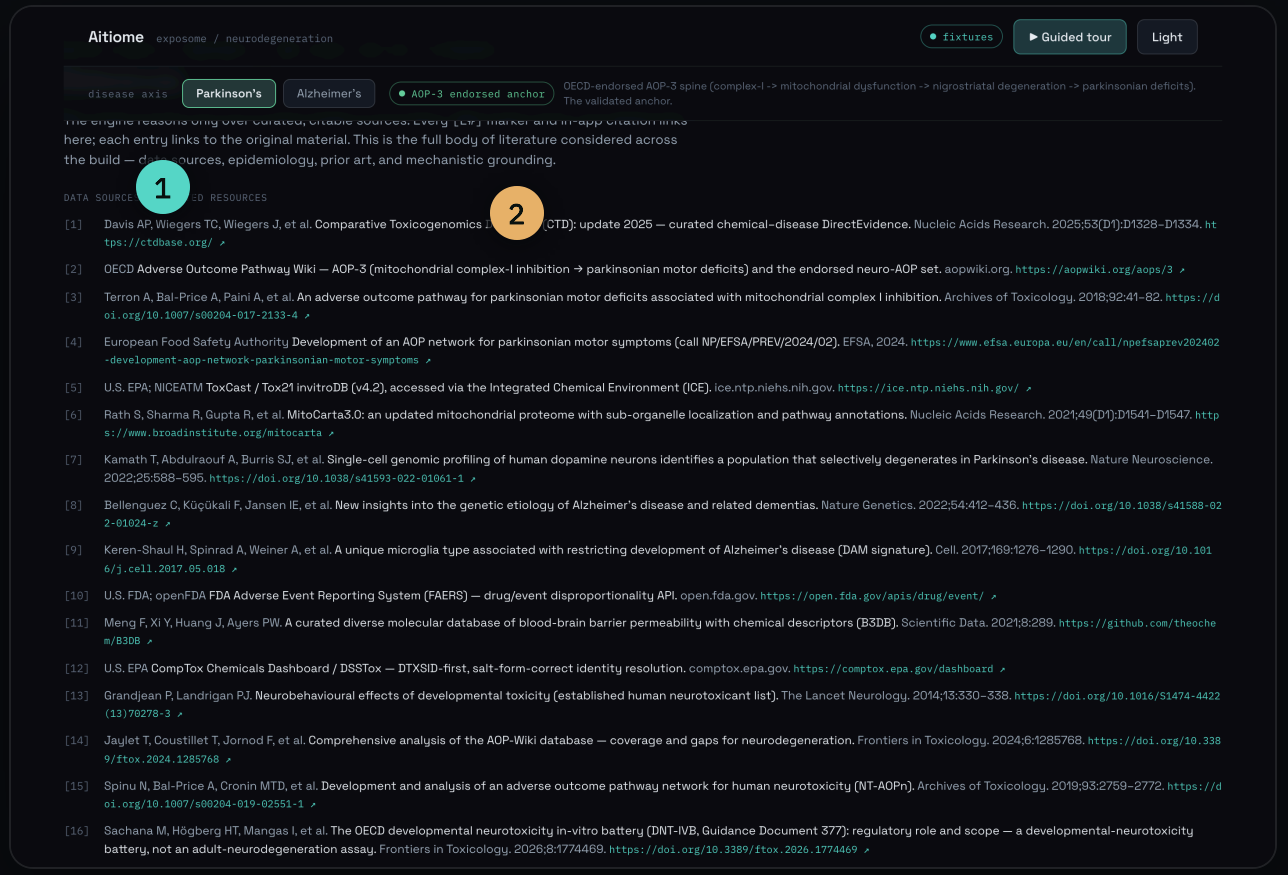
What we deliberately do not use

- **Bioactivity as a discriminator** (ToxCast / Tox21 hitcalls, GenRA, DeepTox) - anti-diagnostic here; ToxCast covers only ~40% of neural-relevant targets (Mack 2024).
- **Circular knowledge graphs** (ComptoxAI, AlzKB, PrimeKG, Hetionet) - their edges are CTD / AOP-derived or drug-keyed, so they would confirm our own inputs.
- **Structure / QSAR & morphology** - general chemical similarity is not neurotoxicity.
- **CTD inferred associations** - inference by study volume (acetaminophen alone has 80 inferred PD links). Only curated DirectEvidence counts.

Anything that smuggles general bioactivity back in, or is circular with our own curation, is disqualified.

AUDITABLE END TO END

Every claim links to its primary source

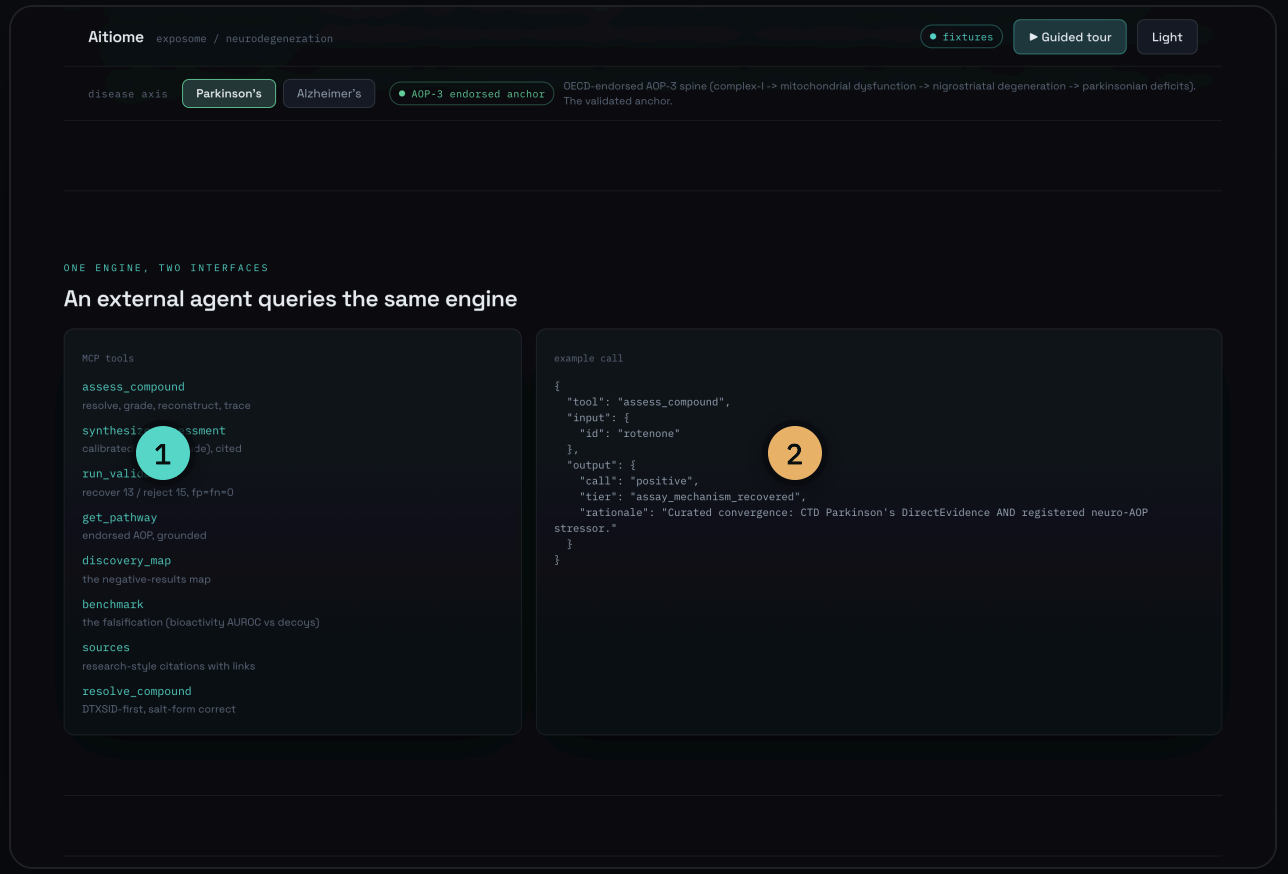


1 Nine primary sources, each a resolvable citation

2 Tagged by role: diagnostic, corroboration, grounding

ONE ENGINE, USABLE BY HUMANS AND AGENTS

An MCP server makes the platform usable by agents, not just people



1 A built-in MCP server exposes eight tools - the exact same engine an agent can drive

2 An agent gets the same graded, cited call a scientist does - human + agentic in one platform

One engine, two interfaces, a model kept out of the decision

versioned contract/ - typed schema + fixtures · core ⊥ viz (the viz runs on fixtures; live is a fixture→flip)

One ~13 MB Go binary serves web + /api + MCP · reproducible: same input, same output · live on fly.io.

HOW THE REASONING WORKS

Six steps per chemical - Claude Code runs the analysis

1
Resolve
Identifier → DTXSID-first, salt-form-correct compound.

→

2
Gate
Deterministic predicate: curated CTD DirectEvidence OR a registered in-scope AOP stressor. Bioactivity never gates.

→

3
Reconstruct
Attach the endorsed AOP (MIE→KE→AO); each edge grounded to a primary source.

→

4
Converge
Assemble the typed evidence strands it reasons over - each with provenance.

→

5
Falsify
Decoys rejected on independent lines; bioactivity shown at/below chance.

→

6
Narrate
Claude (Opus 4.8) via the API writes the calibrated, [E#]-cited synthesis - explains, never decides.

Steps 1-5: the analysis, built with Claude Code

Claude Code wrote the Go engine that runs them - resolution, the curated gate, AOP reconstruction, evidence convergence, decoy falsification. At runtime it's a fixed logical rule over curated evidence: deterministic, auditable, reproducible, no model in the decision path.

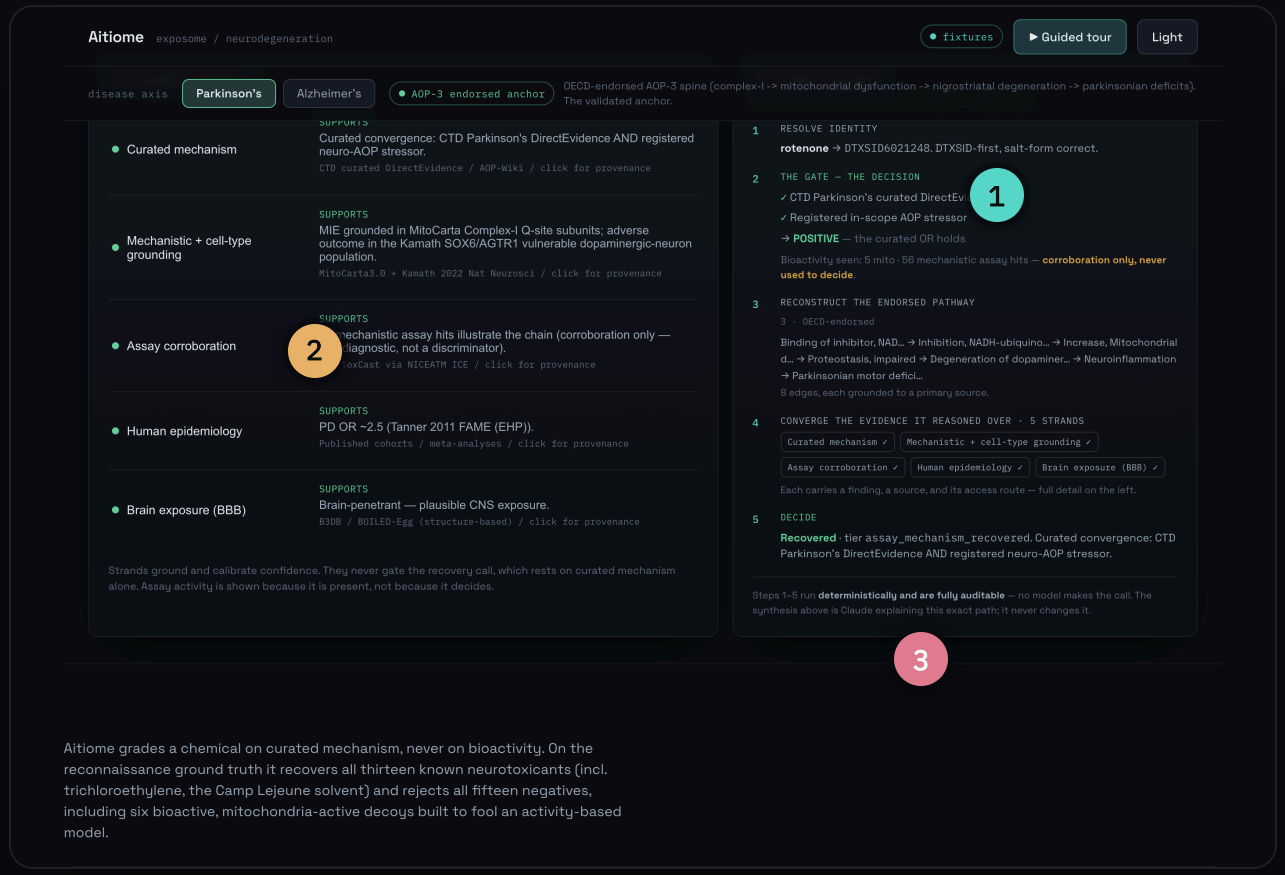
Step 6: Claude via the API (Opus 4.8)

The evidence-reasoner calls Claude to write the calibrated, cited synthesis of the completed assessment. It explains the decision; it can never change the call or the tier.

Claude Code builds the analysis; the Claude API narrates it - and every step emits a trace event you read as "the reasoning path" in the app.

THE REASONING PATH, LIVE IN THE PRODUCT

You can read exactly what it reasoned over, and how it decided



Aitiome grades a chemical on curated mechanism, never on bioactivity. On the reconnaissance ground truth it recovers all thirteen known neurotoxicants (incl. trichloroethylene, the Camp Lejeune solvent) and rejects all fifteen negatives, including six bioactive, mitochondria-active decoys built to fool an activity-based model.

1. The gate: which curated terms fired - and that bioactivity was NOT used to decide
2. The evidence strands it reasoned over - each with a source and access route
3. Steps run deterministically and are auditable - no model makes the call

THE THREE ROLES OF CLAUDE - PRECISELY

Claude builds, verifies, and explains - never decides

Claude Code - builds it

Two parallel streams (engine + visualization) over a versioned contract; the deterministic core, the falsification harness, the hero, the deploy.

Claude API (Opus 4.8) - explains it

The in-app evidence-reasoner: writes the calibrated, [E#]-cited synthesis of a completed assessment. Explains, and never changes the call.

Claude Science - assembles + verifies

Beyond the embedded 27, `make curate` assembles a curated-evidence draft (Claude + web search) verified against primary sources; `/assess-curated` then grades it with the same deterministic gate. An unverified draft stays a hypothesis.

Claude builds, assembles, verifies, and explains - the deterministic gate still makes every call.

Deterministic core, auditable end to end

- One Go binary embeds the curated data; the recovery decision is **deterministic** - no LLM in the decision path.
- Transport-agnostic service, thin HTTP + MCP adapters - a scientist and an external agent query the same tools.
- DTXSID-first, salt-form-correct identity (the paraquat-dichloride trap, tested).
- Every claim links to the original source (CTD, AOP-Wiki, MitoCarta, Kamath, ToxCast/ICE, FAERS, B3DB).
- **Beyond the 27:** Claude Science assembles curated evidence, `POST /assess-curated` grades it - an unverified draft is a hypothesis, never a recovery.
- **We build on standards, not a new formalism:** typed evidence roles (ECO/SEPIO), edge provenance (Biolink/PROV-O), confidence tiers (GRADE), the LLM kept out of the decision path - our contribution is the validation discipline.
- ~13 MB container, live on fly.io. Reproducible: same input, same output.

A SECOND QUESTION WE STUDIED: HOW SHOULD CLAUDE SYNTHESIZE EVIDENCE?

Beyond RAG - Recursive Language Models

RAG - one monolithic pass

A single web-grounded model call reads the sources and writes the answer. Simple and cheap - but it must hold the whole problem in one context, and one slow or failed retrieval sinks the entire pass.

RLM - decompose, then recurse

An Opus **planner** splits the task into ~6 bounded sub-investigations; parallel Sonnet **leaves** each research one; a deterministic merge combines them. Depth-1, per the MIT method.

Decomposition buys coverage, a built-in adversarial critic, and graceful degradation.

Method: Recursive Language Models - Zhang, Kraska & Khattab, MIT CSAIL, arXiv:2512.24601 (2025). Depth fixed at 1; the reproduction study warns deeper recursion inflates cost without accuracy gains.

FOUR SYNTHESIS METHODS, ONE DETERMINISTIC SCORER, SIX CHEMICALS

What we found: RAG vs RAG+ vs RLM vs adversarial RLM

mean per chemical	RAG	RAG+	RLM-1	RLM-ADV
what it is	single pass, plain prompt	single pass, multi-query + counter-evidence prompt	planner + parallel leaves	RLM-1 + adversarial critic
evidence objects	13.5	17.2	14.2	**17.7**
distinct sources	8.2	12.7	8.8	**13.7**
counter-evidence found	1.2	5.7	1.8	**12.7**
cost, 6 chemicals	$2.5	$3.6	$4.0	$7.0

RAG+ and RLM-ADV are the strong variant of each family - RAG+ hardens the single pass with a counter-evidence-seeking prompt; RLM-ADV adds a dedicated critic that hunts refutation. RLM numbers are a **floor**: a slow retrieval window degraded ~2/3 of leaves and they still matched RAG+. All four systems are complete over the six chemicals (24/24 cells).

WHAT THE METHODS STUDY CONCLUDED

Adversarial RLM is the right tool when evidence must be skeptical

The finding

RLM-ADV surfaced **~10x more counter-evidence than RAG and ~2.2x more than the strong RAG+ baseline**, matching RAG+ on yield and source breadth - while running degraded. For a tool built on calibrated honesty, actively hunting *disconfirming* evidence is the property that matters most.

The trade-off, kept honest

RLM-ADV's critic refutes rather than confirms - it recovered 0/6 diagnostic anchors - so it pairs with RLM-1, which builds the case. And decomposition **degrades gracefully**: a stalled leaf loses one sub-question, where a monolithic RAG call loses everything.

The pattern transfers wherever "find the counter-evidence" matters as much as "build the case."

Drug-target validation, pharmacovigilance signal assessment, systematic-review triage, biomarker-disease grading - any life-science synthesis that assembles source-grounded evidence, then stress-tests it.

WHAT WE LEARNED ALONG THE WAY

Challenges, and what they taught us

- **Novelty was the wrong optimization.** Five prior-art scans found the whole design space already published - so we built the validated prototype the field called for, and positioned it accurately.
- **Bioactivity is anti-diagnostic.** The deepest surprise: on our own data it scores worse than chance against the decoys. Activity cannot be the discriminator.
- **Identity correctness is day one.** The paraquat salt-form trap silently returns ”no data” and corrupts everything downstream.
- **Our own test harness caught our own overclaim.** A red-team test we wrote failed, and forced the stronger, evidence-based statement. Ground claims in executable evidence.

THE ADVERSARIAL TEST, TURNED ON OURSELVES

What would falsify Aitiome?

01

A decoy passes the gate

If any adversarial mitochondria-active decoy ever earns a positive call, the curated-diagnostic gate is broken. Six are carried permanently as that control.

02

A blinded benchmark fails

On a larger, blinded neural-specific set, if recovery drops or bioactivity starts to discriminate, the anti-diagnostic claim does not hold.

03

An edge won't ground

If a pathway edge cannot resolve to a primary source, the reconstruction is asserting, not grounding - and we do not ship it.

We state our own failure criteria - the adversarial mindset is the method, not one slide.

THE TAKEAWAYS

What this hackathon produced

Validation that holds

Recovery works (13/13 PD; 12/12 AD) and specificity is proven by falsification - bioactivity at or below chance against adversarial decoys. The discovery limits are mapped, not hidden.

A pipeline that guides the bench

The honest form of discovery: an evidence-weighted triage queue that ranks what to test next, gate-promoted, decoy-controlled, and validated by a held-out backtest. It converges with PROTON, diverges from bioactivity-driven ENRICH.

Calibrated, shipped, live

Built end to end on Claude - which reasons over the evidence but never makes the call - deterministic where it must be, cited to source, live at aitiome.fly.dev with an MCP interface for agents.

And a methods discovery: studying Claude across RAG vs RLM synthesis patterns, the adversarial RLM surfaced ~10x more counter-evidence than RAG - decomposition makes Claude a measurably better skeptic.

Limitations, and the PD / AD roadmap

- A curated benchmark with an adversarial control (27), not a population generalization study.
- Recovery shares curated provenance with the labels - defended by the independent-source ablation and the orthogonal negatives; it is a sanity check, not the headline.
- Anchored on the Parkinson's / mitochondrial route; Alzheimer's is a scaffold extension.
- **Next:** a larger, blinded neural-specific reference set to power the one qualified lead (AUROC 0.72, currently perm-p 0.155).
- **Next:** run the bounded Boltz-2 Q-site benchmark - physics, not annotation or activity.
- **Next:** a time-dated backtest of the candidate queue (rank on pre-dated evidence; show later-curated positives surface) and automated multi-source ingestion.

The discipline transfers where the ingredients exist

The method - reconstruct an endorsed AOP scaffold, gate strictly on curated causal evidence, carry bioactive-but-non-causal decoys as a falsification control - is not specific to neurodegeneration. It transfers to any domain with those same three ingredients. We have **not** demonstrated transfer; this is a falsifiable claim about where the ingredients already exist.

Hepatotoxicity

READIEST ●

mature liver AOPs + dense CTD + DILIrank / ProEuroDILI decoy sets

Skin sensitization

SETTLED SCAFFOLD ●

AOP-40 + OECD TG 497; protein-reactive non-sensitizer decoys

Endocrine disruption

BROADEST SCAFFOLD ●

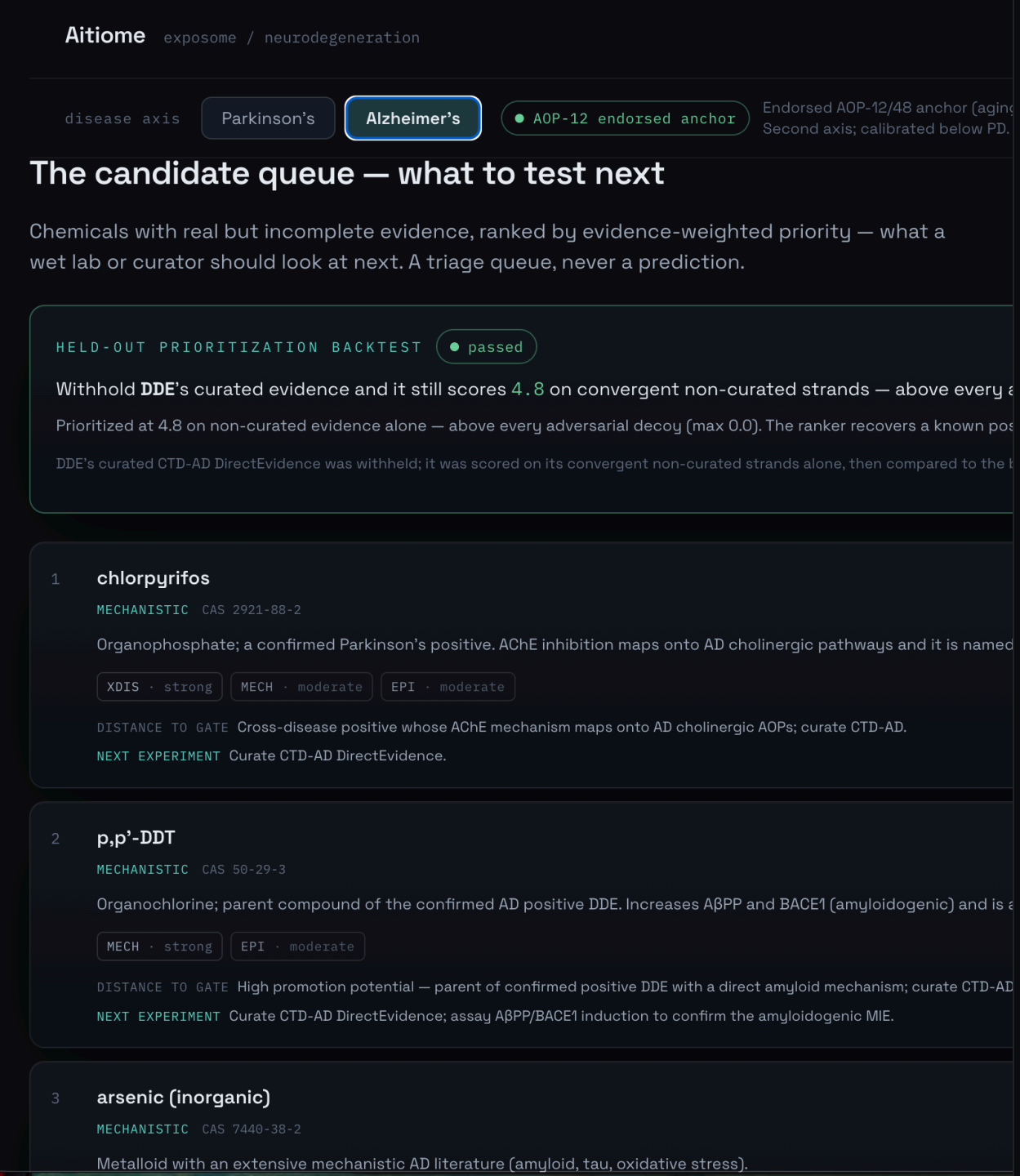
42-AOP EATS network + OECD EDC framework

Cardiotoxicity

CLOSE BEHIND ●

CiPA-anchored; non-torsadogenic hERG blockers as decoys

Honest exclusions: ALS / Huntington's (no scaffold, thin curated evidence - the same line we draw for AD), and carcinogenicity (its gold standard, Key Characteristics, deliberately rejects the single-AOP-chain model).



Honest where the data is thin. Confident where we earned it.

- Validated recovery (13/13 PD, 12/12 AD) with adversarial specificity **proven by falsification** - bioactivity at or below chance against the decoys.
- A candidate pipeline that **guides the bench** (shown live, left) - gate-promoted, decoy-controlled, backtest-validated.
- A methods finding: **adversarial RLM surfaced ~10x more counter-evidence than RAG** (~2.2x vs RAG+) - decomposition makes Claude a measurably better skeptic.

Try it live at **aitiome.fly.dev** / github.com/jonradoff/aitiome