How do you use Knowledge Graphs to drive decision making?

What we've learned encoding scientific decisions in a graph database, told through an example from our client work.

PREPARED FOR	DATE	SPEAKER
GraphTalk Boston	2026-10-15	Hamza Farooq, CTO BioBox

How did you get here this morning?

A quick show of hands.



Drove

Your own car



Walked

On foot from nearby



Car service

Taxi, Uber, Lyft



Can't remember

No hands, hopefully

One small decision, made a few hours ago.

Would you remember why?

Same question, asked later. The decision fades slowly; the reasons go first.

	Today	*In a week*	*In a month*	*In a few years*
THE DECISION	**Drove**	**Drove**	**Drove, I think**	**?**
THE WHY *What influenced it*	**Time** **Cost** **Convenience** **A chance to rest**	**Time** Cost Convenience A chance to rest	Time, maybe?	—

The decision outlives its reasons.

In drug discovery, the why is what matters

Consequential calls, made constantly. The reasoning behind them ends up scattered.

CALLS THAT MATTER

Advance a target
Drop a target
Choose an indication
In-license an asset
Set the TPP

Years of work and real money ride on each one.

WHERE THE WHY ENDS UP

Scientists' presentations
Rebuilt for each meeting, then filed away

Executive summaries
The conclusion, without the path to it

Excel sheets
tracker_v14_FINAL.xlsx

Fragmented systems
ELN, LIMS, warehouses, shared drives

Meeting transcripts
Said once, recorded, rarely found again

Email threads
Reply-all archaeology

And in people's heads, including the people who have since moved on.

THE GAP

The decision survives. Its why doesn't, so it gets re-argued and recomputed.

An example: an ALS program and a new VP

Drawn from work with our clients. Most of you will know a version of it.

COMPANY	Halden Bio: mid-size biotech, neurodegeneration focus
PROGRAM	ALS target prioritization, 178-gene field
THE DECISION	Which targets advance at the next portfolio review
THE NEW VP	Dana, VP of Discovery, three weeks in
HER QUESTION	**“What are we doing, and why?”**

COMPUTATIONAL BIOLOGY

Priya

A 60-slide network analysis. Rigorous, and unreadable outside her team.

TOP 3: SCN1A, TARGET X, GPR-7

MEDICINAL CHEMISTRY

Marcus

A notebook and a decade of pattern-matching. Usually right; can't always say why.

TOP 3: GPR-7, MAP-3, LRR-2

DATA TEAM

The dashboard

Scores everywhere. Lineage nowhere.

TOP 3: MAP-3, SCN1A, TRP-9

THE PREVIOUS VP

A folder

target_tracker_v14_FINAL.xlsx, and a note that says “ask Marcus”.

TOP 3: TARGET X, TRP-9, LRR-2

Four top threes, barely overlapping. Each confident; none can say where the others' numbers came from.

“Why did we deprioritise Target X?”

The question every review eventually asks. Dana asks all four.

PRIYA	“I think there was a screen. Let me find the deck.”
MARCUS	“That was a chemistry call. It's in my notebook somewhere.”
DATA TEAM	“The Teams channel has it.” → https://teams.microsoft.com/… (404)
TRACKER	Excel tracker cell F14 turned red between v12 and v13. No note.

The decision happened. The evidence existed. The connection was never stored.

WHAT DANA NEEDS TO ANSWER

1. **What is the state of every decision?**
   Live, parked or dropped, and who owns each call.
2. **Why is any target ranked where it is?**
   All the way down to the analysis behind the evidence.
3. **What changed since the last review?**
   What moved, by how much, and which evidence moved it.

THE GAP — **Four answers. Zero provenance. No new data needed; the problem is structural.**

A decision is a structure, and it has altitude

Everyone is right. They just work at different levels, with nothing connecting them.

DECISION	Which targets advance. Dana lives here.
FRAMEWORK	Scientific priorities and scientific criteria. The team's judgment, written down.
EVIDENCE	Target X: “central in the motor neuron network” 0.82 · from NX-14
DEDICATED ANALYSES	Co-expression networks, pathway models, screen analyses. Priya lives here.
RAW DATA	Answer ALS scRNA-seq, Project MinE, internal screens, ELN entries.

THE MISSING LINKS

Dana needs the top layer. Priya works in the fourth. With nothing connecting them, every question turns into a meeting.

BIOBOX **Connect the layers: Halden's framework, encoded against Halden's data in a graph database, so every layer links to the one below it.**

A decision without its why is just a vote nobody remembers taking.

One foundation, many portfolio decisions

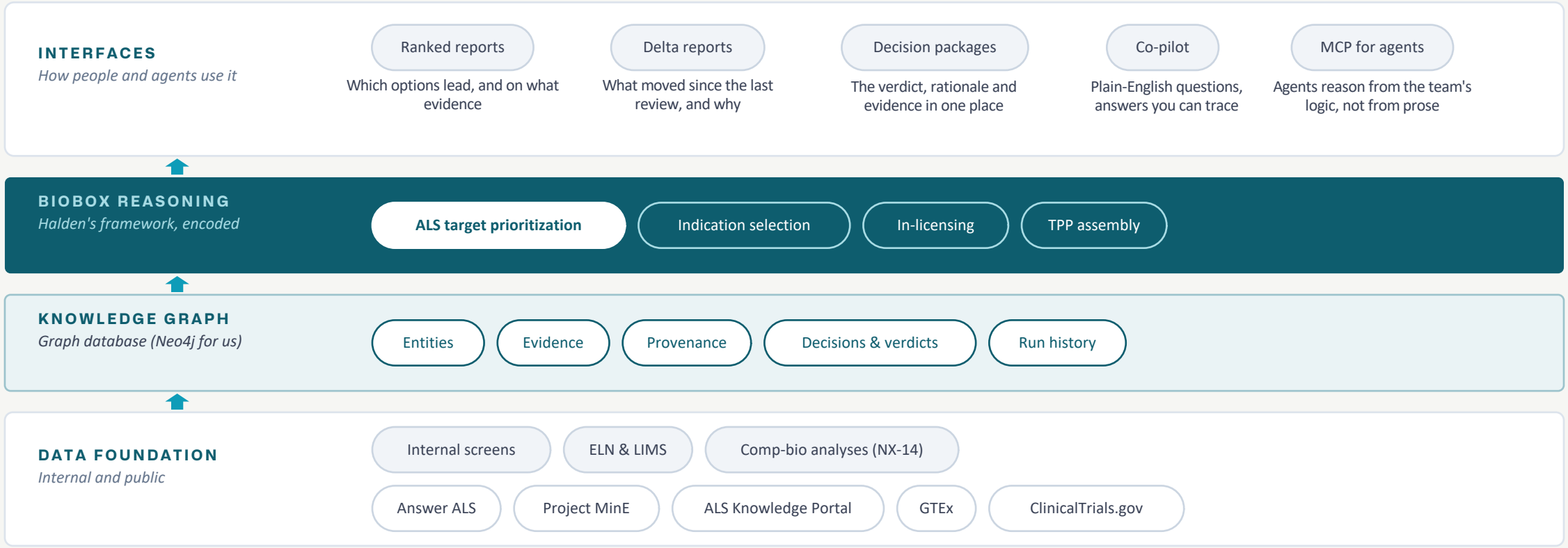
Every portfolio decision breaks down the same way. Laid out along the lifecycle.

DISCOVERY	TRANSLATIONAL	DEVELOPMENT	PORTFOLIO & BD
TARGET PRIORITIZATION *Which ALS targets should advance at the next portfolio review?*	**INDICATION SELECTION** *Which indications should our IL-23 asset expand into next?*	**TPP ASSEMBLY** *Does this profile beat standard of care on the criteria that matter?*	**IN-LICENSING SCREEN** *Which partner assets fit our immunology strategy and TPP?*
TARGET EVALUATION *Which surface antigens should we pursue for an ADC in bladder cancer?*	**BIOMARKER SELECTION** *Which biomarkers best stratify likely responders for this mechanism?*	**ASSET REPURPOSING** *We have a VEGFA inhibitor in RCC. Which other indications would it suit?*	**PORTFOLIO REVIEW** *What should we do next, and why has that changed?*

ONE FOUNDATION — **One decision graph · linked to the team's data where it lives · each team's logic, encoded as a reasoning system**

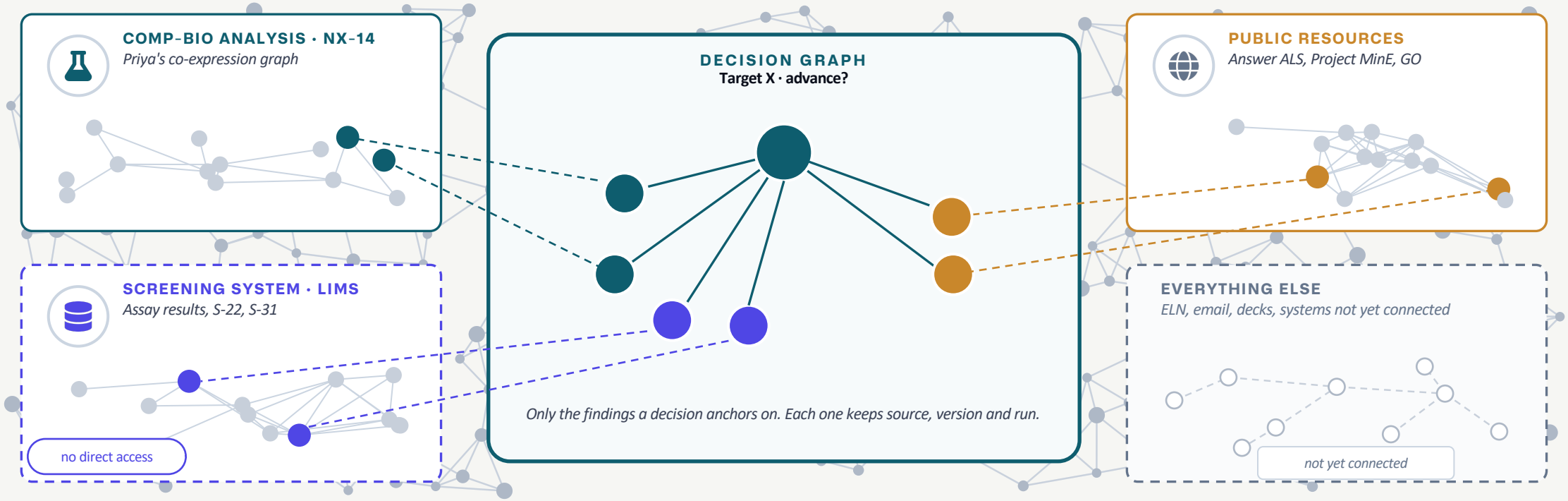
What sits underneath

Halden's data and Halden's framework, with one graph underneath. In theory.



One graph in theory. Many graphs in practice.

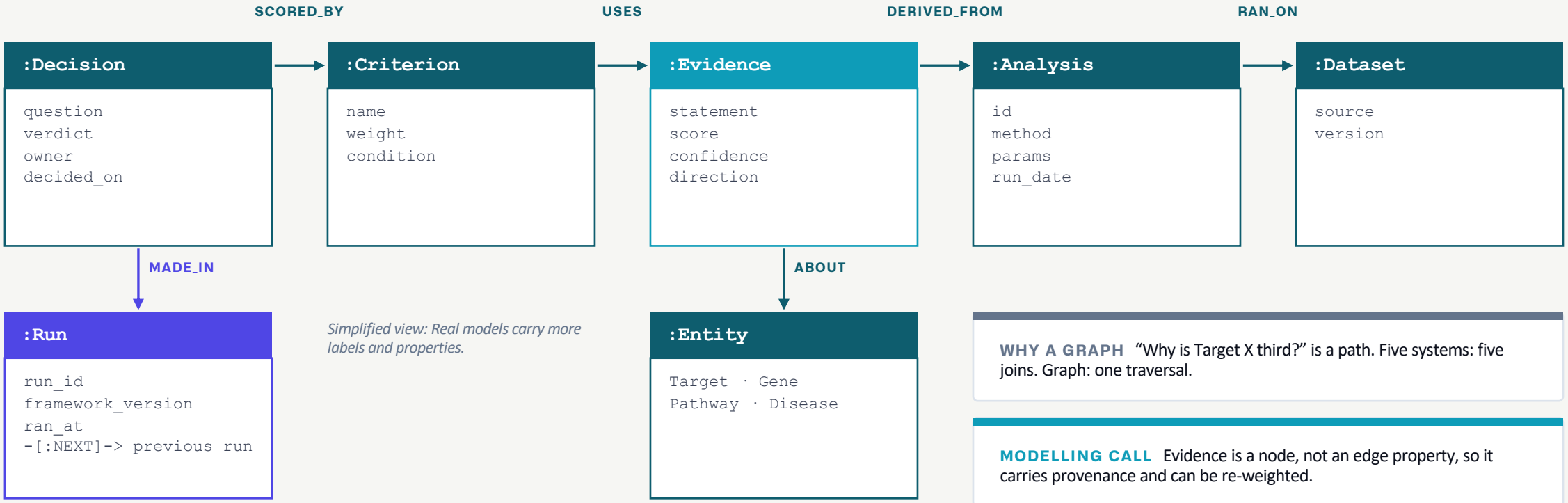
The decision graph holds the findings it anchors on, with provenance back to where each one lives.



THE RULE **Extract the finding. Keep the provenance. The decision knows where the data lives, even when it can't reach it.**

The question is a path, so the model is a graph

A simplified schema. Entities, evidence and provenance are all nodes, and so are runs.



Simplified view: Real models carry more labels and properties.

WHY A GRAPH "Why is Target X third?" is a path. Five systems: five joins. Graph: one traversal.

MODELLING CALL Evidence is a node, not an edge property, so it carries provenance and can be re-weighted.

Same graph, three ways in

Cypher for Priya. Plain English for Dana. Tools for Halden's agents.

CYPHER

```
MATCH (t:Target {name:'Target X'})
      -[:SCORED_BY]->(c:Criterion)
      -[:USES]->(e:Evidence)
      -[:DERIVED_FROM]->(a:Analysis)
RETURN c.name, c.score,
       e.statement, e.confidence,
       a.id, a.run_date
ORDER BY c.score DESC
```

DANA TYPES	“Why is Target X ranked third?”
ENTITY RESOLUTION	“Target X” → node id. “ranked” → scoring criteria.
TEMPLATED CYPHER	Safe fragments assembled from a plan, not freeform generation.
ANSWER	Criterion-by-criterion breakdown for Dana. Priya clicks through to NX-14.

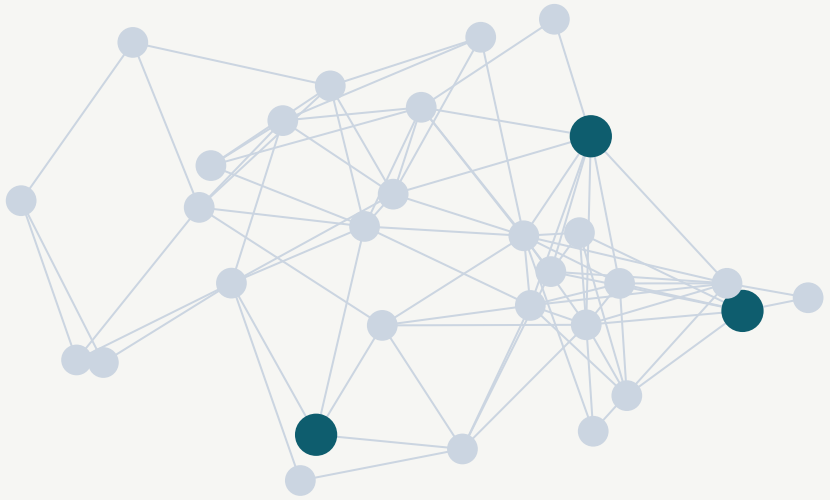
FOR AGENTS · BIOBOX MCP

`list_reasoning_systems` · `run_reasoning_system`
`get_evidence` · `get_graph_schema`

Ranked results, the team's criteria and the evidence trail flow back to the agent. Authenticated per client: not everyone builds, everyone can query.

Deep analyses don't get replaced. They get connected.

Priya's motor neuron network analysis becomes one evidence line, with a pointer back.



CO-EXPRESSION NETWORK · NX-14

Thousands of nodes. A 60-slide deck. Priya's domain.

ONE EVIDENCE LINE

Target X is central in the motor neuron disease network

ASSOCIATION SCORE	0.82
COUNTS TOWARD	Expression → the composite score
SOURCE	NX-14 · Answer ALS scRNA-seq
RUN	2026-08-19

↳ open full analysis

One conclusion, one evidence line, one pointer back. The depth is still there for anyone who wants it.

Four criteria, one composite score

Each criterion is scored on its own evidence. The composite is built from those contributors, and every one is traceable.

TARGET X · COMPOSITE

0.74



How much each criterion counts is set by Halden's scientists, not by us.

TOP EVIDENCE LINES

Central in motor neuron co-expression network

NX-14 · Answer ALS scRNA-seq · 0.82

Rare variants enriched in ALS cases

Project MinE · ALS Knowledge Portal · 0.91

Druggable pocket, two tool compounds

ChemDB-2 · 0.55

Independent scores roll up into one composite, so no single criterion can hide the others.The criteria, and how much each counts, belongs to the team

Where the criteria and their balance come from

The hardest part isn't the graph. It's writing down judgment that has never been written down.

TODAY

Criteria are applied, but never written down

Every call already weighs criteria
Genetics against tractability against safety, in someone's head.

The balance feels qualitative
“Genetics matters more.” More by how much?

Every team weighs differently
And the balance shifts from meeting to meeting.

Nothing is written down
So nothing can be checked, reused or improved.

WHAT WE DO INSTEAD: CALIBRATE AGAINST CALLS THE TEAM ALREADY TRUSTS

1 Elicit
Draft the criteria with the scientists: what makes you advance a target?

2 Encode
Criteria are encoded, with a starting balance and penalties.

3 Calibrate
Run on targets the team already expects near the top. See where they land.

4 Adjust
Rebalance until the known calls land right. Log why.

Repeat until the calls the team already trusts land where they expect. Then the calibrated criteria become the template for the next decision.

CALIBRATION LOG · HALDEN'S TARGET CALL

CRITERION	CHANGE	WHY
Genetics	**Raised**	Familial genes ranked too low
Expression	**Lowered**	Rewarded broad expression
Tractability	Unchanged	Unchanged
Safety	Unchanged	Unchanged

ANCHOR CHECK

#6 of 178

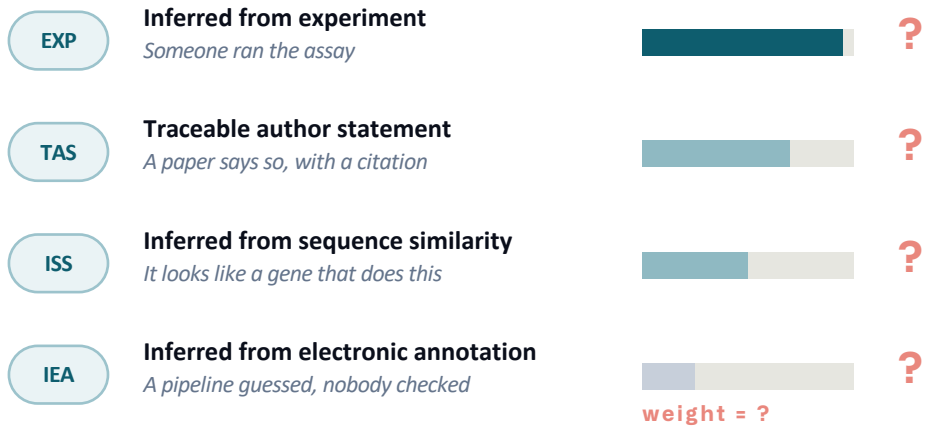
SOD1, with an approved antisense therapy, lands where the team expects it.

Quantitative is easy to weight. Qualitative isn't.

The weight on an edge is obvious when the edge carries a number. Most evidence doesn't.

EXAMPLE · GENE ONTOLOGY EVIDENCE CODES

Same annotation, four ways of knowing it. How much more is an experiment worth than an electronic annotation?



HOW WE TUNE IT TODAY · THE FEEDBACK LOOP

1. **Start defensible**
   Order the codes the way the field would. Pick starting weights.
2. **Run the known candidates**
   Targets the team already trusts, like SOD1 in ALS.
3. **Check the positives came out**
   Did they land where the team expects? If not, the weight is wrong, not the target.
4. **Adjust and log why**
   Every change is recorded with its reason. Then go again.

What you're capturing is what your data told you, or at least what you saw it say that matched up.

AN OPEN ASK — **If you have a more structured way to weight qualitative evidence, find me afterwards. I'd love to dig into it.**

Back at Halden: a screen finishes two weeks before the review

Instead of v15 of the tracker, Dana got the delta since the last review.

TARGET	LAST REVIEW	NOW	DELTA	DRIVER
GPR-7	5	2	**+3**	New screen hit. Tractability +0.30
SCN1A	4	19	**−15**	Expression module: two changes since last review
SCN2A	6	22	**−16**	Expression module: two changes since last review
Target X	3	3	—	S-31 hits. Tractability +0.16 (already rank 3)
MAP-3	8	6	**+2**	Literature evidence added

A field of 178 ALS genes. Every row answers "what moved, by how much, and why" before anyone asks.



LIVING CHECKPOINTS

Nobody re-ran this by hand. Codified criteria are re-checked every time data is added or removed. What's newly met or newly failed raises a flag.

The SCN family fell in two stages

One delta since the last review. Stepping through the runs in between split it into two causes.

SCN FAMILY · RANK BY RUN

Last review → mid-cycle run → now

SCN1A	4	→	9	→ **19**
SCN2A	6	→	12	→ **22**
SCN3A	11	→	17	→ **27**
SCN8A	14	→	20	→ **31**

Without the run history: one unexplained drop of 15.

STAGE 1 · LAST REVIEW → MID-CYCLE

Evidence removed upstream

An expression source was refreshed. Evidence lines supporting the SCN family dropped out of the expression criterion.

TEAM CALL

Accepted.
The refreshed data is better.

STAGE 2 · MID-CYCLE → NOW

Analysis re-run with new parameters

NX-14 was re-run with a stricter cutoff. The remaining SCN evidence fell below it.

TEAM CALL

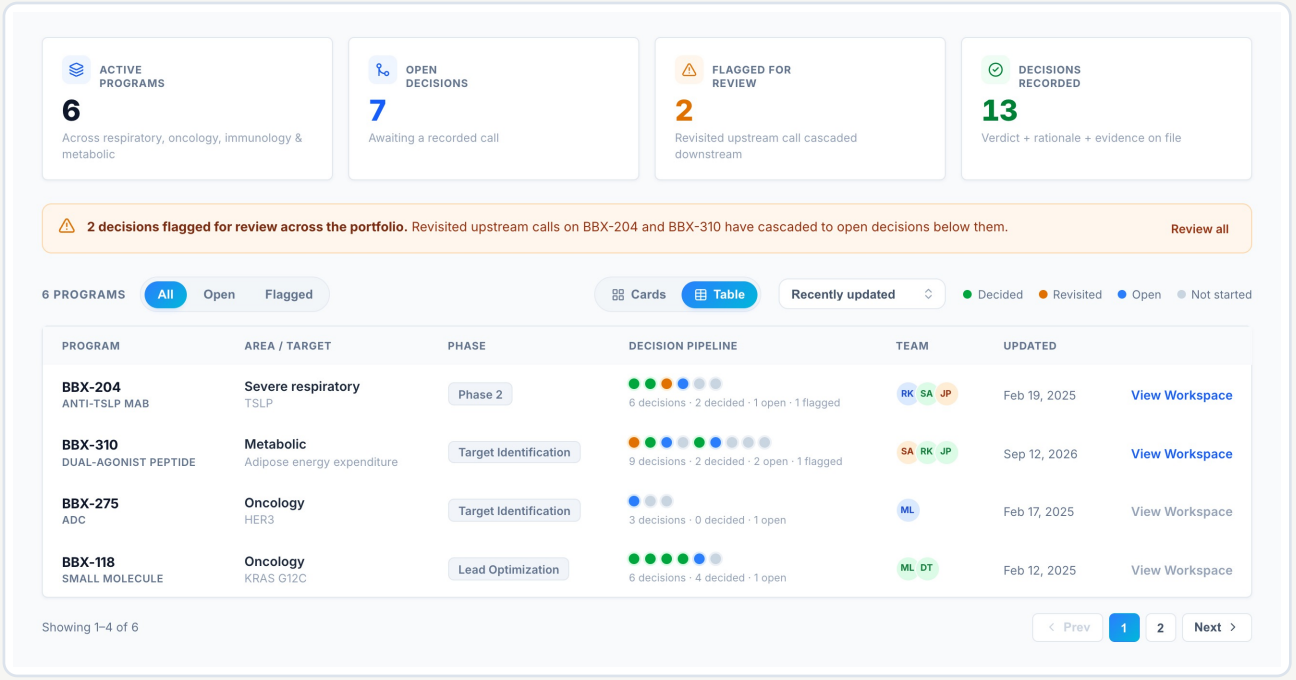
Challenged.
Priya reviewed the cutoff before the review.

THE ANSWER **Nothing happened to the biology. Two things happened to the pipeline, and the graph could tell them apart.**

Walking into the review

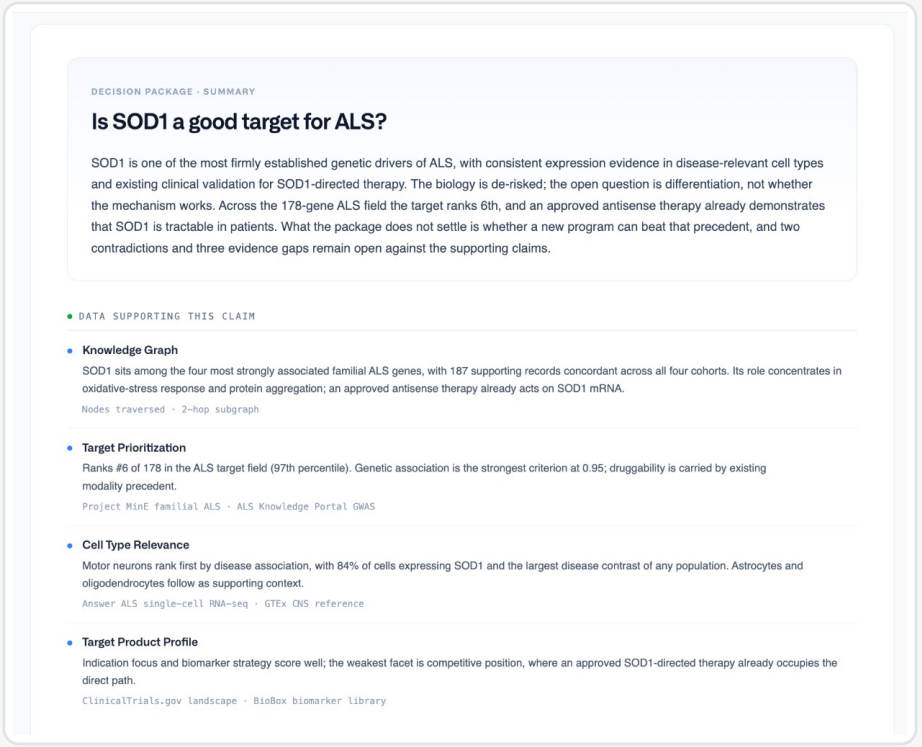
Dana's first question, answered: the state of every decision, with the rationale one click away.

DECISION COMMAND CENTER



The flag banner is the living checkpoints at work: revisit an upstream call and every decision below it is flagged.

DECISION PACKAGE · ALS



Every call opens to its verdict, claim and supporting data.

“Why did we deprioritise Target X?”

In the review, someone asks the question from her first week.

The decision.
The evidence.
The change.
One traversal.

Fifteen seconds, with provenance.

DECISION

Deprioritised 2026-03-14 · owner: Marcus
Tractability fell below threshold after the ChemDB-2 refresh.

EVIDENCE IT RESTED ON

Druggable pocket · ChemDB-2 · 0.55 (was 0.78)
No tool compound passed selectivity · Screen S-22

WHAT CHANGED SINCE

New screen S-31: two selective hits · Tractability back to 0.71
Recommend re-opening.

THE GAP, CLOSED

One answer. Full provenance. Nobody left the room to check Slack.

What changes for a team like Halden's

Four things Dana would tell the next VP.

	BEFORE	AFTER
ONBOARDING	A quarter of meetings	**A week reading the graph**
REVIEWS	Argue about where the numbers came from	**Argue about the science**
DEEP ANALYSES	Sit in 60-slide decks	**Feed decisions as evidence lines, with the author's name on them**
AUDIT TRAIL	Built by hand before diligence	**Exists by default**

Codified decisions make agents practical

Three ways an agent can meet a decision. One discloses progressively: the decision first, evidence on demand.

READS EVERYTHING

Exhausts its context

Pulls every paper, report and evidence line. Fills the window, burns tokens, and can answer differently on every run.

CONTEXT USED

- × Uses the team's criteria
- × Knows what changed
- × Stays within budget

TAKES IT AT FACE VALUE

Repeats the verdict

Finds “Target X: deprioritised” in a summary and passes it on. No reasons, and no idea whether it's still true.

CONTEXT USED

- × Uses the team's criteria
- × Knows what changed
- ✓ Stays within budget

PROGRESSIVE DISCLOSURE

Discloses progressively

Starts from the verdict, criteria, owner and what changed. Pulls only the evidence that moved, and cites the run.

CONTEXT USED

- ✓ Uses the team's criteria
- ✓ Knows what changed
- ✓ Stays within budget

THE POINT

The decision graph is an agent's starting point, not its last resort.

Prose tells an agent what you think. Structure tells it how.

The same criterion, two ways to hand it to an agent.

AS PROSE · A SYSTEM PROMPT

> *"We favour targets with strong genetic support and good expression in disease-relevant cells. Tractability matters, but less than genetics. Penalise anything with safety concerns. Use your judgment."*

- ✕ **"Strong" means what?** Two agents, two readings. The same agent, two runs.
- ✕ **How do the criteria combine?** Nothing says how genetics trades off against safety.
- ✕ **What counts as evidence?** Not which sources, not which analyses, not which version.

AS STRUCTURE · IN THE GRAPH



- ✓ **A weight, not an adjective.** How much genetics counts against safety is explicit, not implied.
- ✓ **Evidence lines, each with a source.** The agent traverses to them; it doesn't guess what counts.
- ✓ **A run it was calibrated on.** Same input, same answer. And it can show its work.

FOR YOUR AI ROLLOUT **Codified decisions give agents guidance and structure that a prose description can't.**

Conference paper: presenting at NeurIPS 2026, Sydney.

What we're still working out

Where we've landed so far. Better answers welcome.

How do you version a decision?

Not the data: the decision. When the weights change, is it the same decision?

WHERE WE'VE LANDED

Every decision points at the run and framework version it was made on. New weights mean a new run; the old call stays replayable.

Where does human sign-off live when an agent proposes a re-rank?

WHERE WE'VE LANDED

Agents can propose. A named owner records the verdict and the rationale.

Who owns the weights when two teams share a criterion?

Say, translational and BD.

WHERE WE'VE LANDED

Honestly, still more a governance conversation than a technical one.

When should an upstream change re-open a decision?

Automatically, or just raise a flag?

WHERE WE'VE LANDED

Today we flag; we don't re-open. The owner decides.

Better answers, or better questions? Find me after the panel.

Three takeaways

Three things to take back with you.

OWNERSHIP

If you can't replay a decision, you don't own it.

Ask your team to show you the chain behind any one ranking. If the answer is a meeting, that's the gap.

THE MODEL

Evidence, provenance and runs are nodes.

Every evidence line points down to the analysis it abstracts. Every decision points at the run it was made on.

THE ANALYSES

Deep analyses and the decision graph aren't rivals.

The graph is how the analyses reach the people who act on them.

A decision without its why is just a vote nobody remembers taking.

What should we do next, and why has that changed?

Hamza Farooq · CTO, BioBox · biobox.io

QUESTIONS