A GENERAL INTELLIGENCE · A NEW WAY TO FIND CURES

Discovering cancer's hidden rules from raw data.

Heather is Aphorion's general intelligence. Pointed at raw cancer data, with no labels, features, or hints, it rediscovered known biology and redrew the cell's internal map. It proves two of its core claims, and it opens a new way to find the gene combinations that could one day help solve cancer.

GENERAL INTELLIGENCE · THE FUTURE OF CANCER DISCOVERY

18,000 × 1,100
GENES × CANCER CELL LINES

58 → 8
MACHINES → SYSTEMS RECOVERED

~1,900
COMBINATION LEADS PROPOSED

CPU
NO GPU · NO NEURAL NET

The discovery

> Given only a table of raw cancer measurements (about 18,000 genes switched off across 1,100 cancer cell lines, the public DepMap screen), with no labels, no expert features, and no hints, Heather rediscovered known cancer biology, redrew the cell's internal map, and proposed roughly 1,900 gene-combination leads to investigate.

It pulled out textbook machinery it was never told about, and two of its highest-ranked drug-combination predictions are real targets already in human trials. Heather reached all of this blind, before anyone looked anything up, which is how we know what it saw was real structure and not noise. This report is two things at once: evidence that a general intelligence can read a disease's own rules from raw data, and a first look at what that could mean for finding cures.

Heather is AGI

Heather is Aphorion's general intelligence. It perceives raw information directly, discovers the hidden laws inside it, recalls by content, and knows when something is genuinely new, abstaining rather than guessing. It learns from a single example and does not forget. The same machine that reads a cancer screen reads a physics trajectory or a market. It is a general instrument, and cancer is where its first proof and its largest stakes happen to coincide.

RAW PERCEPTION | LAW DISCOVERY | ASSOCIATIVE RECALL | ABSTAINS ON NOVELTY | ONE-SHOT LEARNING | NO FORGETTING

UNDER THE HOOD

No GPUs. No neural networks. No encoders, no decoders, no training. Heather runs on a laptop, on a CPU. This report states the capabilities and proves two of them; the substrate that achieves them is out of scope here.

The two AGI claims this proves

A general intelligence rests on abilities most systems fake with hand-built machinery. Two of them Heather has, and the cancer result proves both, because in cancer the right answers are already known and each claim can be graded against reality.

> CLAIM 1
>
> ### Raw perception
>
> An intelligence must perceive raw information directly, with no person first engineering the features for it. Most AI skips this: a vision model is handed pixels only after human-designed preprocessing, a language model is handed tokens rather than raw text. Heather takes the raw stream, here the unprocessed gene-effect table, and finds what matters in it with nothing supplied.

> CLAIM 2
>
> ### Hidden-law discovery
>
> Storing information is not understanding it. The leap is to recover the rules that generated the observations: the quantities that stay fixed and the things that move together. From the raw cancer numbers Heather recovered the laws of the cell. It is the same operation that, given raw motion, recovers a conservation law in physics.

The future of cancer research

Cancer is not solved because the search is vast and slow. Thousands of genes, hundreds of tumour contexts, an astronomical number of possible combinations, and every promising pair costs months at the bench. Progress is gated less by ideas than by how quickly hypotheses can be generated, triaged, and trusted.

A general intelligence that reads raw data and finds its laws changes that arithmetic. Point it at a screen (any tumour, any measurement type) and in minutes it returns the couplings worth testing, ranked and traceable back to the data, with the obvious confounders already stripped out. The two axes it found here are already in the clinic; the other ~1,900 are waiting. Point it at data no one has mined yet and it could surface combinations no one has seen.

The future of cancer research may look less like a sharper microscope and more like an instrument that reads the disease's own rules and says where to strike. Dataset by dataset, that may be how cancer is finally solved.

The data and the signal

The DepMap screen knocks out each of about 18,000 genes, one at a time, across 1,100 human cancer cell lines, and records how much each knockout slows or kills each line. The result is a matrix: a row for every cell line, a column for every gene, and in each cell a number, the fitness cost of losing that gene there. A gene's entire description, to Heather, is its column, the vector of costs across all 1,100 lines. No gene name carries meaning, no pathway is annotated, no feature is computed.

Two genes are **co-dependent** when their columns move together: lines that need gene A also need gene B, and lines indifferent to A are indifferent to B. Co-dependence is the footprint of shared function. The two genes sit in the same molecular machine, the same pathway, or in a synthetic-lethal relationship where the loss of one forces reliance on the other. Heather measures it as the correlation of the two raw columns, computed for every pair of genes; the rest of the procedure refines that quantity into trustworthy links.

The discovery procedure

A bare correlation is cheap and mostly spurious. A co-dependency is reported only when it satisfies every condition below. Each is a property the link must hold, not a step we perform, and each removes a distinct class of false positive.

CONDITION	WHAT IT REQUIRES	WHAT IT REMOVES
Coupling	raw correlation ≥ 0.30	weak, incidental pairs
Unrecorded	absent from the STRING interaction database	links already known, isolating candidates
Reproducible	holds on a held-out 70 / 30 split of the cell lines	flukes of the particular sample
Context-invariant	within-tissue correlation keeps ≥ 55% of its strength	tissue-of-origin artifacts
Mutual	each gene sits in the other's top neighbours	one-sided hub effects
Cross-module	the two genes' neighbour sets do not overlap	trivial within-complex pairs
Active	both genes are expressed in the lines carrying the link	couplings on genes a cell never switches on

The conditions run in this order so the cheapest filters cut the search before the expensive ones. The last is mechanistic rather than statistical: a cell cannot depend on a gene it does not transcribe, so a link between two silent genes is measurement noise and is discarded, however strong the correlation. Crucially, "absent from the database" flags candidates but does not certify them. The database is incomplete, so a link can be unrecorded yet real; final novelty is settled only by a literature review, then by experiment.

Discover blind, then confirm

Heather is run with no biological input and returns ranked co-dependencies. Only afterward is each top result checked against the published literature. A link the literature confirms is evidence that the procedure reads real biology; a link it has not recorded is a candidate for the bench. The order is the safeguard: the answer is found before any annotation is consulted, so agreement with the textbooks cannot be circular. Every headline result in this report was produced in that order.

The tissue control, and its calibration

The largest source of false co-dependency is tissue of origin. Two genes can correlate across the whole panel simply because both are essential in, say, melanoma, with no functional link between them. The control removes this by measuring the correlation *within* each tissue and requiring the link to survive. Its trustworthiness is established by running it on two pairs whose answers are already known.

PAIR	RAW R	WITHIN-TISSUE	VERDICT
BRAF · MAPK1 (same pathway)	+0.679	+0.573	survives ✓ (real)
BRAF · MITF (melanoma lineage)	+0.347	+0.164	collapses ✗ (artifact)

A known functional pair, the cancer gene BRAF and the signalling kinase MAPK1 it acts through, keeps most of its strength. A known lineage artifact, BRAF and the melanocyte factor MITF, loses most of it. The control passes the true link and rejects the false one, which is why its verdicts on the novel links can be trusted. The two other quantitative gates work the same way: the held-out split asks whether a link measured on one set of cell lines still holds on a set it was never shown, and the expression gate asks whether the link lives in cells that actually run both genes.

What the procedure is not

The method does not predict therapeutic synergy. Co-dependency establishes that two genes are needed together; whether co-targeting them outperforms targeting one is a separate, experimental question. Nor does the method invent biology absent from its input: on a screen the field has mined for years, a correct procedure recovers established structure, which is precisely what makes the recovery a clean test of the two claims rather than a novelty contest.

The two confirmed axes

Of the links Heather ranked highest, two were checked exhaustively against the literature. Both are real, both are druggable, and both were reached blind. They are not presented as new biology; they are the proof that the perception is veridical, and the template for the leads still to be tested.

METTL14 · MRE11: RNA tagging meets DNA repair

Heather's strongest novel signal places METTL14, the partner subunit of the m6A RNA-methylation writer, with the DNA-repair nuclease MRE11. The internal check is immediate: METTL14's single nearest neighbour is METTL3, the enzyme it is physically bound to (correlation +0.58), exactly what the biology demands; its next neighbours are MRE11 and the replication factors ORC3 and MCM6. The link is unusually robust. It reproduces on held-out lines (train +0.52, test +0.54), and under the tissue control it does not weaken at all (within-tissue +0.534 against a raw +0.523), cleaner than the BRAF–MAPK1 positive control. It is strongest in brain, breast and nerve-sheath tumours, across more than 80 lines, a pan-cancer link rather than a single lineage. The literature confirms the axis: METTL3 is recruited to DNA double-strand breaks and is required for homologous-recombination repair; m6A modulates the response to PARP inhibitors; and a parallel methyltransferase, METTL16, was shown to be synthetic-lethal with PARP inhibition through MRE11. The drug is in the clinic, the METTL3 inhibitor STC-15 is in Phase 2 trials for sarcoma, which turns a METTL3-plus-PARP combination into a concrete, testable hypothesis.

Field validation

The METTL14 axis was shown to researchers in medicine and pharmaceutical R&D. The reading from the bench matches what Heather found from the numbers alone, and turns it into a concrete drug strategy.

> *"The correlations make sense. They tell us two parts are involved — METTL14 and METTL3 — where the former depends on the latter to synergistically deactivate a critical step in DNA repair, which is mediated by RNA. We already have a lead compound, STC-15, that inhibits METTL3, which essentially renders METTL14 of no effect because it needs its counterpart. That is why the report shows their activities, with respect to this specific pathway, so highly positively correlated. So it makes sense to select one target — METTL3, an enzyme in this case — and attack it with a molecule like STC-15 to limit its activity and automatically inhibit METTL14, just because we know they are highly correlated. It's very relevant medically. We also have PARP inhibitors that could be added to the mix to increase the lethal load. It's good and practical, and could help speed up computational drug discovery."*
>
> MORRIS WAITHAKA · PHARMACEUTICAL R&D

The clinical reading also explains the strength of the signal. Because a single molecule acting on METTL3 collapses the activity of METTL14 as well — the partner cannot function without it — the two genes' dependencies track each other almost exactly, which is the correlation Heather ranked first. One target, one drug, two genes silenced; a PARP inhibitor stacks a second hit on the same DNA-repair vulnerability. This is one expert reading, to be confirmed with others, but it places the finding squarely in practical, druggable territory.

BRD9 · KAT2A: remodeller meets acetyltransferase

The second flagship places BRD9, a subunit of the ncBAF chromatin-remodelling complex, with KAT2A, the catalytic enzyme of the SAGA acetyltransferase complex. BRD9 is a validated dependency: synovial sarcoma, driven by the SS18–SSX fusion, becomes addicted to ncBAF, and BRD9 degraders are already in trials. The novel element is the partner. The link survives the tissue control (within-tissue +0.31 from a raw +0.38, the profile of a real functional coupling), and it is reciprocal: asked from the other side, KAT2A's own top co-dependants include BRD9 and the BAF subunit SMARCD1. It is strongest in soft-tissue lineages, which include synovial sarcoma, exactly where the primary target is being drugged. The specific BRD9–KAT2A coupling is not in the interaction databases, with one caveat: both are general transcription coactivators, so a non-specific "transcription-addicted cells need both" explanation is not yet excluded.

The cell's architecture, recovered

The two confirmed axes are single links. The deeper result is structural. Apply the same correlation operation to Heather's own output, treating each recovered machine as a new object, and a hierarchy appears: genes group into molecular machines, and machines group into systems.

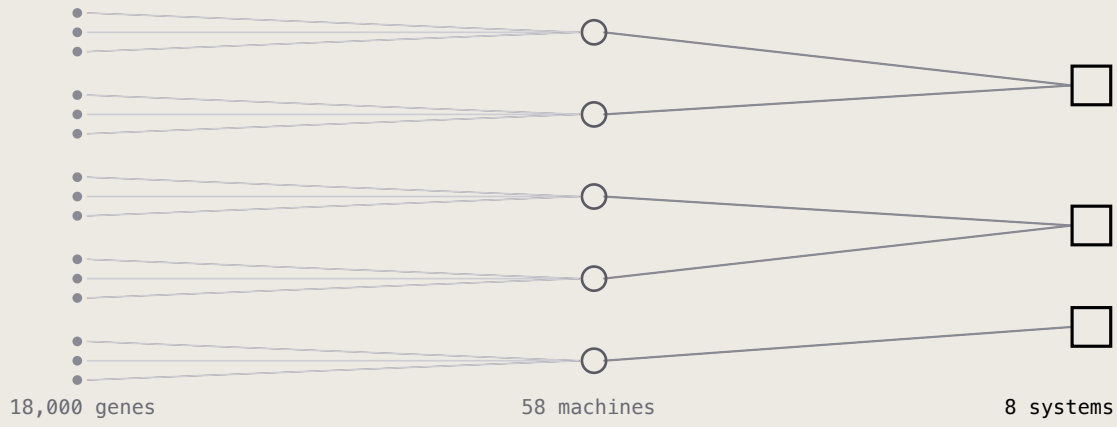


FIG 1. **THE CELL'S HIERARCHY, RECOVERED FROM RAW NUMBERS.** ONE OPERATION, APPLIED TO GENES AND THEN TO ITS OWN OUTPUT, GROUPS 18,000 GENES INTO 58 MOLECULAR MACHINES AND THOSE INTO 8 FUNCTIONAL SYSTEMS.

The 58 machines are recognisable protein complexes (SWI/SNF, SAGA, the Fanconi-anemia repair group, ATP synthase, the purine-synthesis enzymes), recovered with no annotation. The 8 systems are the cell's job description: making energy, copying and repairing DNA, building proteins, processing genetic messages, holding the cell's shape, signalling, moving cargo, and dividing. Run the operation a third time, on the machines themselves, and it finds laws *between* complexes: SAGA cooperating with the Mediator complex, and the two opposing systems that switch genes on and off (MLL and PRC2) turning out to be co-essential, because the cell needs the balance between them. No single rule dominates the whole: the largest pattern, roughly "how fast the cell grows," explains only about a tenth of the variation. What results is close to the organisational chart of a living cell, assembled from raw numbers by one repeated operation.

LEADS The ~1,900 candidates

Beyond the confirmed axes, about 1,900 gene pairs cleared every condition yet are absent from the databases. Only a handful have been hand-checked, and database-absence overstates true novelty, so these are leads to investigate, not discoveries: the queue a wet lab, or a fresh dataset, would work through next.

The honest catch

"Needed together" is not "a cure." A reported pair means two genes matter to the same cells. Whether hitting both does better than hitting one is a laboratory question that data alone cannot answer, and every lead here carries that proviso. The vision of solving cancer is exactly that, a vision; this report proves a capability, not a treatment.

The output is bounded by the input. The test ran on public data the whole field has studied for years, so most of what surfaced was already known. That is what makes the result a clean proof of the two claims, and it also caps how novel the output can be on this particular dataset. The genuinely new biology, and the cures, would come from data not yet mined.

The novelty lives in the data; the capability lives in Heather. The instrument is the durable asset. The same two abilities proven here, raw perception and hidden-law discovery, carry to any field that produces data to perceive.

Bottom line

> Heather perceives raw information and discovers the laws inside it, with no hints and no neural networks. On cancer it rediscovered real, drugged biology and redrew the cell's map, proving a general intelligence and, in the same stroke, a new way to search for cures. The proof is here; the cures are what the next datasets are for.

Reproduce: leaderboard.py (gene-pair leads), meta_laws.py (the cell map). Input: public DepMap 24Q4 release (CRISPRGeneEffect, Model, OmicsExpression) and the STRING interaction database; all public.

Confirmations: Wainberg et al., Nat Genet 2021 (co-essential modules) · Michel et al., Nat Cell Biol 2018 (ncBAF / BRD9 / synovial sarcoma) · eLife 2021 and Nature Cancer 2022 (m6A, MRE11, DNA-repair, PARP) · STORM Therapeutics, ASCO 2024 (METTL3 inhibitor STC-15 in trials).

A P H O R I O N L A B S