Evidence-Driven Target Discovery: Reconstructing Disease-State Transitions with Mechanistic Knowledge Graphs

How connected biological evidence enables more confident target prioritization and AI-assisted hypothesis generation in Acute Myeloid Leukemia.

WHITEPAPER

elucidata

Contents

Executive Summary

A target discovery program today can draw on genomic, proteomic, single-cell, and functional-screen data at real scale. Open Targets alone catalogues 8.1 million target-disease associations; UK Biobank, TCGA, DepMap, GTEx, and the Human Cell Atlas have each substantially expanded the evidence base available to a discovery program. However, only about 1 in 50 discovery programs ultimately reaches the market.

The gap is not due to missing information but rather due to data being spread across the full text of the publications, supplementary files, tables, figures and clinical readouts. One way to overcome this is by using out-of-the-box knowledge graphs that aggregate this information. But the challenge there is that these KGs capture generic biological information but don't contain context about your therapeutic area.

This whitepaper describes an approach to building disease-specific knowledge graphs - a base layer of curated public evidence, augmented with context mined from publications, databases, and proprietary data that we made queryable in plain language through AI assistants via the Model Context Protocol (MCP). Applied to Acute Myeloid Leukemia, the graph reconstructed the disease's underlying mechanism in under half an hour, converging on **revumenib, an approved menin inhibitor** already active in NPM1-mutant disease. The same query flagged two further points on that axis: DOT1L, which already has an investigational compound in AML trials, and MEIS1, a transcription factor with no compound in the graph at all - a tractability gap, not a shortfall in the evidence, and consistent with how hard transcription factors are to drug directly.

AML affects roughly 20,000 people in the US each year and carries a five-year survival rate near 30%. In close to a third of adult cases, it is driven by a recurrent mutation in NPM1 that stalls hematopoietic progenitors mid-differentiation instead of letting them mature. The path reconstructed here runs from that mutation through HOXA9, and separately through MEIS1 and KMT2A into the menin complex, converging on the same revumenib-targeted axis. Each link had been published independently, in different papers and formats, disconnected from one another until queried together.

The case study demonstrates how integrating connected, typed evidence enables efficient reconstruction of mechanistic relationships, facilitating the generation of testable hypotheses

This framework can be applied to any indication for surfacing drug-repurposing opportunities through shared biology and identify novel targets, even where data is scarce.

Introduction: The Structural Limits of Current Target Discovery

Open Targets translates literature and functional genomics into 8.1 million target-disease associations. DepMap contributes genetic dependency evidence from functional knockout screens across hundreds of cancer cell lines. GTEx supplies the tissue-specific expression context needed to interpret a genetic association in the tissue where a disease actually manifests. Together, these platforms represent a real expansion of the human genetic evidence base a program can draw on.

The prioritization funnel has not changed shape. Discovery programs still start from an initial candidate pool on the order of 20,000 genes - overwhelmingly protein-coding, though therapeutic targets increasingly extend to non-coding RNAs and other biotypes as well and have to narrow that pool to two to five experimental candidates. That process still takes years and substantial capital, and most programs that enter it do not produce an approved therapy. Only about 1 in 50 discovery programs ultimately reaches the market.

The challenge is no longer identifying candidate targets. It is selecting the few supported by evidence strong enough to justify therapeutic investment.

Three structural problems sit underneath that funnel.

The traditional drug discovery bottleneck is manual synthesis

Evidence relevant to a specific hypothesis is spread across full-text papers, tables, figures, and proprietary datasets, and most of that biological context is never structured or made queryable. The result is that scientists lose real time to manually curating data that already exists somewhere in the literature.

Commodity knowledge graphs answer a general biological question -

Off-the-shelf knowledge graphs aggregate public datasets and literature-derived relationships at scale. Although they may include evidence scores, confidence measures, or predefined ranking systems, they are generally not constructed around the biology of a specific disease program or a project-specific set of prioritization criteria.

A query returns a connection between a gene and a disease; it does not indicate whether that connection is more biologically significant than the dozen other connections returned alongside it, because significance depends on criteria that are use-case specific and simply not encoded in a generic graph.

Co-occurrence is not equivalent to biological association-

Open Targets is among the most widely used platforms for target discovery. It scores gene-disease relationships primarily through literature co-occurrence: a gene and a disease named in the same sentence, sometimes the same paragraph. The literature evidence score reflects factors such as the location and frequency of the co-occurrence within the publication, rather than the semantic strength or direction of the biological claim.

A sentence stating there is no evidence of an association between a gene and a disease, and a sentence stating the gene is associated with that disease, name the same two entities and get scored as comparably strong evidence. This is because the scoring has no mechanism for reading the claim itself, only for detecting that the two names appeared together.

Open Targets is simply the most visible example of a much broader category: any graph built primarily on co-occurrence inherits this same blind spot, regardless of how many sources it aggregates or how large its evidence base grows.

The practical consequence is identical no matter which of these platforms sits upstream: reading the primary literature, reconciling it against experimental data, and constructing a defensible mechanistic hypothesis remains a manual step.

Approach: An Evidence- Centric Infrastructure for target discovery

Biology should be modeled around a specific therapeutic question, not general data aggregation. Even in a disease with substantial published literature, the mechanisms that matter - genetic dependencies, chromatin regulation, functional screens - stay scattered until they are deliberately connected, weighed, and scored against that specific disease context. That connective work is what Elucidata's evidence infrastructure is built to do.

The infrastructure has four components - Scout, Xtract, Polly KG, and Lens. Polly KG can be connected to any AI assistant through Polly MCP and each can be used as a standalone tool; together, they form a pipeline to move from a disease question to a narrowed and high-quality shortlist of targets.

Finding and curating the right evidence

Polly Scout operates over a continuously curated corpus of **8 million full-text publications** and **360,000 public biomedical datasets.** Rather than returning every record that happens to mention a disease term, it identifies the datasets and publications that are actually relevant to the specific disease question being asked.

Polly Xtract then extracts and maps the relevant content out of those sources - literature, supplementary tables, dataset metadata onto a common multimodal schema, preparing it for integration into the graph rather than leaving it as a disconnected pile of source material.

Structuring the evidence as a layered knowledge graph

Polly Knowledge Graph holds that curated evidence in three layers, each adding a different kind of specificity on top of the last.

Layer	What it holds	Illustrative sources
Base KG	Pathway, disease, ontology, and molecular relationships. Available from day one, for any disease.	NCBI, UniProt, Reactome, ChEMBL, Open Targets
Contextual	Disease-, tissue-, phenotype and species-specific evidence identified for the exact question at hand. (We add context-specific data according to the program's specifications.)	GEO, PubMed, TCGA, CCLE, DepMap, Cancer Hallmark, cBioPortal
Proprietary	A program's own internal data, plus a jointly built custom scoring framework with domain scientists.	Internal experimental data; program-specific prioritization criteria

12.44M
nodes across 13 node types

40.39M
edges across 183 edge types

20+
curated public sources

8M+
full-text papers

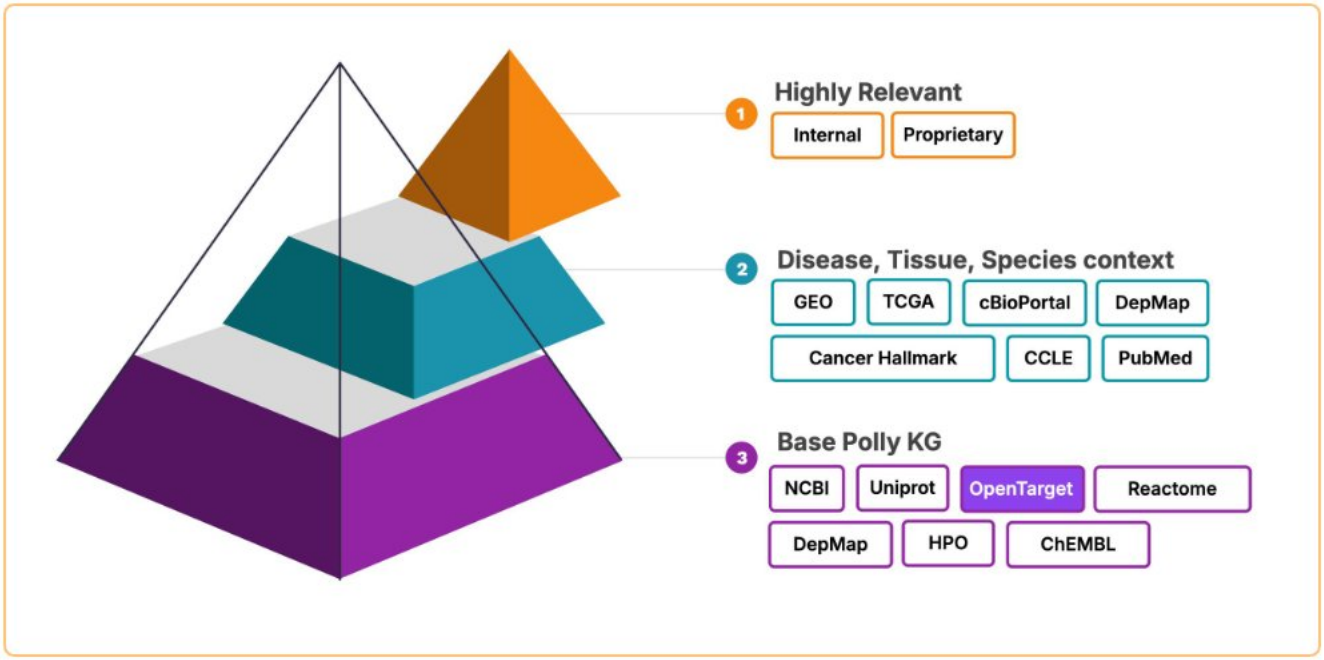


Figure 1.Polly KG architecture: data sources, graph contents , and capabilities. The system supports natural-language querying from day one.

DATA SOURCES ·

what goes in

- NCBI · UniProt · ChEMBL
- Open Targets · Reactome
- GO · UBERON · MONDO
- GEO
- Full-text papers via literature mining
- Custom in-house / proprietary data

POLLY KG ·

what it contains

- Genes · diseases · drugs
- Pathways · cell types
- Phenotypes
- Directed, typed relationships
- Multi-source confidence scoring
- Ontology-aligned entity resolution
- Dynamically updated

CAPABILITIES ·

what you can do

- Natural language querying (NLQ)
- Multi-hop traversal across entities
- Hypothesis generation
- Linkage + association prediction
- Tractability & druggability lookup
- Functional-perturbation (CRISPR) queries
- Custom scoring framework
- Cross-species reasoning

Two capabilities that move beyond retrieval

1. CUSTOM SCORING

- Polly knowledge graph prioritizes targets and biomarkers based on disease-specific criteria.
- The same piece of evidence can be decisive for one program and irrelevant to another, depending entirely on what that program is trying to prioritize. Polly KG does not pre-score its base layer for that reason.
- Scoring is developed jointly with a client's domain scientists and applied only once their actual priorities are understood - weighed across dimensions such as tissue relevance, species, pathway centrality, molecular tractability, and how consistently a finding replicates across independent studies.

2. EDGE (LINK) PREDICTION

- A model trained on the graph's existing structure - which genes share complexes, which pathways recur together, which entities cluster by evidence type - can infer that two entities are very likely connected even when no single publication states the connection directly.
- The model predicts novel relationships between entities, identifies previously unreported gene-disease or disease-drug associations and surfaces high-confidence hypotheses for experimental validation. These relationships are added to the knowledge graph.
- Predicted edges are flagged distinctly from directly cited ones - they are semantically supported hypotheses worth testing, not established findings.

Enriching Open Targets: From Flat Associations to Graded Evidence

Open Targets remains foundational to Polly KG's base layer - it is one of the most systematically maintained sources of gene-disease evidence available, and Polly KG is built to extend it. The limitation described earlier is structural: A single association score does not define association/relationship type. An established finding, a probable one, and a speculative one all can collapse into the same number based on gene-disease mention.

Take a real evidence sentence, drawn from PMID 36499755: "GOLM1 and FAM49B may be used as markers for the assessment of HNSCC prognosis and as potential targets for HNSCC treatment."

Open Targets records this as one fact - FAM49B associated with HNSCC, score 1.0.

Polly KG parses the same sentence into three separate, independently graded claims:

- Expression altered in disease - established, FAM49B is differentially expressed in HNSCC tissue.
- Prognostic biomarker - probable, Proposed as a marker for HNSCC prognosis.
- Therapeutic target - speculative, Named as a potential target, with certainty graded down from the hedge "potential."

The distinction matters because these three claims carry different weight in a prioritization decision, even though all three originate from the same sentence. Separating what is established from what is inferred, and scoring each independently, is what the contextual and proprietary layers described above are built to add on top of Open Targets and every other base source in the graph.

Make Evidence AI-Accessible

Polly MCP connects the graph to Claude, GPT, Gemini, or an enterprise model through a single standard connector. Every layer described above, schema, evidence-type scoring, multi-hop traversal, tractability, functional perturbation, and protein-complex membership, becomes reachable from one natural-language question, submitted directly to the AI assistant a research team already uses.

This matters more than it might first appear. Most biomedical knowledge graphs require the person asking the question to already know the schema: which relationship types exist, how disease terms map to ontologies, how to write a correct query in a graph language such as Cypher. That requirement alone restricts direct access to a small population of computational specialists, no matter how good the underlying evidence is. MCP changes what is required to ask, not what the graph contains. A scientist submits a biological question in plain language, and gets back a scoped, cited result: a ranked list, a mechanistic path, a directed subgraph, whatever the question calls for.

It is worth being equally clear about what this does not change. The reliability of any answer still depends on the completeness and evidential quality of the graph underneath it, a dependency that predates MCP and applies to any system built on curated data. And a conversational interface does not substitute for the statistical judgment a trained bioinformatician brings to validating a hypothesis. It lowers the barrier to asking the question, not the standard required to interpret what comes back. MCP delivers the evidence faster; the judgment needed to weigh it stays with the scientist.

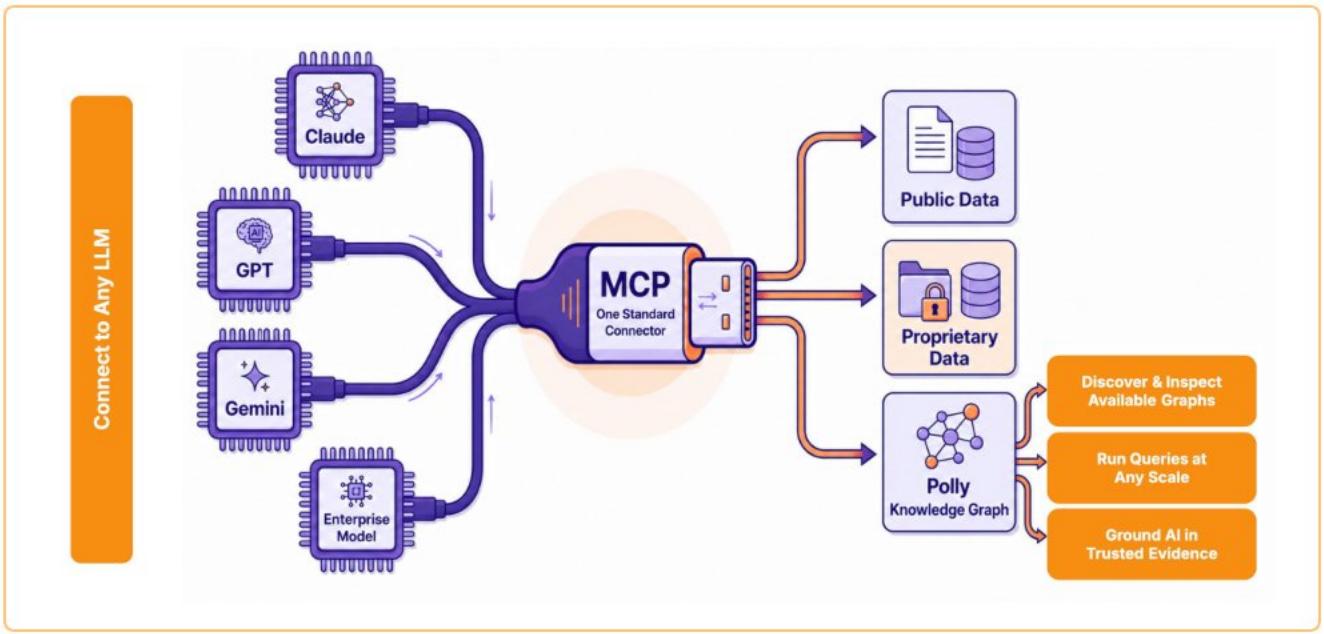


Figure 2.Polly KG MCP as a single standard connector: any LLM in, public data, proprietary data, and Polly KG out, with discovery, querying, and grounding handled the same way regardless of which assistant is asking.

From graph to shortlist

The pipeline above is designed to move a search space of roughly 20,000 genes to about 20 putative targets through graph-based exploration and scoring.

Distilling 20 Putative Targets into 2 Validated Pre-Clinical Targets

Polly Lens runs a structured, cited literature review against a defined question set for each target in the context of disease and produces a documented evidence package to support the final selection.

Every part of that pipeline - the graph, the scoring, the literature review is reachable directly from a natural-language question through Polly MCP.

Case Study: Reconstructing the NPM1-Mutant AML Differentiation-Arrest Axis

AML's biology is extensively published: the graph was given only a disease name, with no pre-selected gene list, to see whether it could reconstruct the underlying mechanism on its own. A disease this well characterized also sets a high bar for the result. The mechanistic path the graph produces either holds together against decades of published biology, or it does not.

The disease

Acute Myeloid Leukemia is an aggressive hematological malignancy in which hematopoietic stem and progenitor cells fail to differentiate normally, allowing immature leukemic blasts to accumulate instead of maturing into functional blood cells. It affects roughly 20,000 people in the US each year, with a five-year relative survival rate near 30%. NPM1 mutations are present in close to a third of adult AML cases - the single most common recurrent mutation in the disease.

Discovery question

Could a knowledge graph, queried entirely in natural language, reconstruct the mechanistic path from a disease name to a druggable target - without a scientist first knowing which genes, which edges, or which evidence layer held the answer?

What querying looks like

In this case study , Polly KG is connected to Claude through MCP. Every question below was asked directly in Claude, in plain English, with no query language or prior knowledge of the graph's schema. Claude passes the question to Polly KG and returns its answer, with supporting evidence, inside the same conversation.

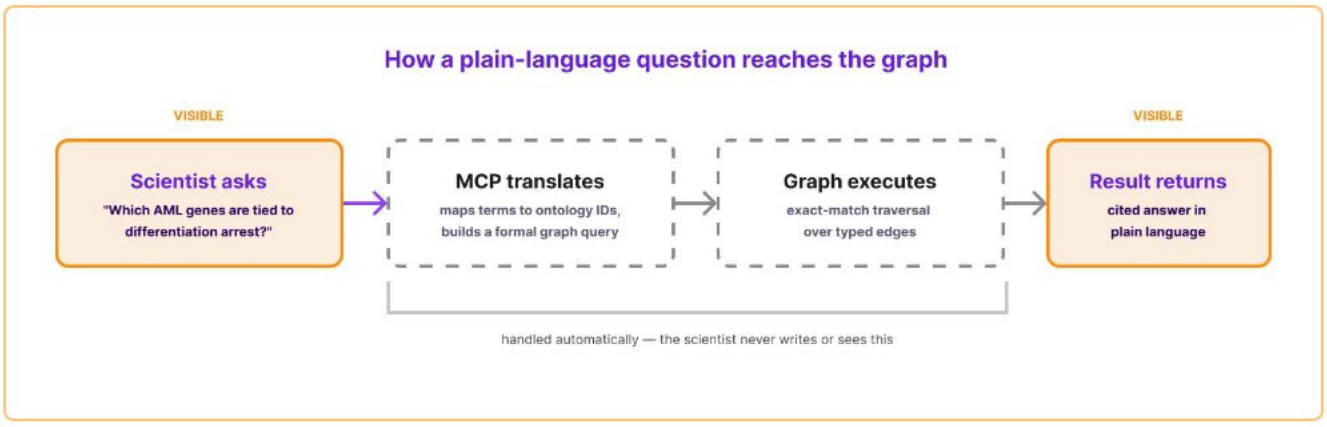


Figure 3. Data Flow- a plain-language question becomes a formal graph query and comes back as a cited answer.

Everything from this point on - every gene list, every table, every score, is Claude's response, grounded in Polly KG.

Mapping the graph before mapping the disease

Before asking anything about AML specifically, the first question put to the graph was simply what nodes and edges it currently held - no node names, no relationship types were supplied in advance:

Node type	Count
gene	8,998,935
drug	2,921,148
protein	227,286
cell_line	167,186
disease	28,296
biological_process	24,428
pathway	23,498
phenotype	19,389
tissue	14,971
molecular_function	10,056
protein_complex	5,638
cellular_component	4,076

Totals: approximately 12.44 million nodes across 13 node types, and approximately 40.39 million edges across 183 edge types.

Edge type	Count
has_essentiality_in	16,615,290
interacts_with_gene	13,441,463
interacts_with_protein	5,902,672
co-occurs_with_gene_in_literature	2,089,112
has_phenotype_model_gene	726,091
linked_to_go	285,778
has_genetic_association_with	166,567
has_phenotype	142,347
has_differential_expression_of	143,234
encodes	100,963

A follow-up question about a single edge type, has_differential_expression_of, returned not just the node types it connects but the properties riding on that edge: log2 fold change, p-value, tissue context. Understanding the shape of the evidence took two natural-language questions and no prior knowledge of the schema.

Ranking AML-associated genes by evidence type

With the shape of the graph established, the first real question: which genes carry the strongest AML evidence, ranked by how many independent kinds of evidence support them rather than by a single score? One translation step happens first. MCP maps the plain-language term "AML" to Acute Myeloid Leukemia and then to its formal disease-ontology identifier, because the graph itself matches on exact identifiers, not free text. Once that resolution is done, the graph returns:

Rank	Gene	Evidence types	Avg score	Key evidence types
01	CEBPA	11	0.95	curated causative gene, clinically validated, driver mutation, diagnostic, dependency
02	KRAS	11	0.87	biomarker, driver mutation, somatic pathogenic variant, phenotype model
03	DNMT3A	10	0.95	biomarker, drug target, driver mutation, burden association
04	JAK2	9	0.0.87	biomarker, drug target, driver mutation, somatic variant
05	FLT3	9	0.86	biomarker, drug target, dependency, driver mutation
06	RUNX1	8	0.90	somatic mutation, phenotype model, dependency, diagnostic
07	KIT	8	0.84	biomarker, drug target, driver mutation
08	TP53	8	0.84	biomarker, drug target, phenotype model, driver mutation
09	IDH1	7	0.94	biomarker, drug target, driver mutation
10	ETV6	7	0.78	diagnostic, phenotype model, somatic variant

Also notable: ASXL1 (average score 1.0 across 4 evidence types - somatic mutation, burden association, driver mutation, co-occurrence), GATA2 (0.984 across 5 types), and SRSF2 (0.983 across 6 types).

CEBPA, DNMT3A, KRAS, FLT3, and JAK2 lead this list. None of them wins on a single high score and many disease-gene edges cap out at 1.0, so ties are common. They lead because each one is backed by several kinds of evidence at once: clinical validation, known drug-target status, driver-mutation data, and diagnostic or biomarker support, all pointing the same way.

FLT3, JAK2, IDH1, and IDH2 carry extra clinical weight because they are already actionable drug targets in AML.

WHY THIS MATTERS

NPM1 and KMT2A are textbook AML genes, but they rank lower here - not because the evidence is weak, but because most of their support sits at the level of AML subtypes, not the general AML disease node. Querying a subtype directly next brings it out.

Tissue specificity as a filter, not a requirement

Evidence strength says nothing about where in the body a gene is active. A gene can be strongly tied to AML and still be doing routine work in every tissue, which changes how interesting it is as a target. So the next question layered in tissue expression, drawn from TPM-based data:

Gene	Tissue(s) with clearly elevated expression
CEBPA	skin - suprapubic (139), lower leg (127); adipose (101, 78); liver (78–40)
KRAS	pancreas (41), brain (37), stomach (31), lung (29) - broadly distributed
RUNX1	stomach (8.0), lung (7.6), intestine (4.6) - low absolute expression everywhere
TP53	ovary (215), spleen (98), blood (77)
IDH2	muscle (3,822 - very high), heart (1,943), skeletal muscle (461)
NPM1	blood (963 - highest of all genes/tissues in this set), connective tissue (860), ovary (794)

NPM1 and IDH2 show the strongest tissue signal - NPM1 in blood, matching its hematopoietic role, and IDH2 in muscle. Others, like RUNX1 and KRAS, show low or broadly distributed expression with no dominant tissue. Only CEBPA and IDH2 carry a formal “tissue-enhanced” tag in the graph's own scoring.

WHY THIS MATTERS

RUNX1 is the reminder here: a gene can be central to AML biology without needing tissue enrichment to prove it. Tissue specificity narrows the picture; it doesn't gatekeep it.

The differentiation-arrest program

AML's defining failure is that blood cells stop maturing partway through development. Any gene that matters therapeutically has to connect to that failure somehow. So the next question asked which of the ranked genes are linked to differentiation arrest and to leukemic stem-cell maintenance. The graph answered using standard Gene Ontology (GO) terms already attached to each gene, not a proprietary classification of its own:

Gene	GO annotation	AML evidence (strongest)
HOXA9	negative regulation of myeloid differentiation	somatic mutation (0.75), literature (1.0)
MEIS1	negative regulation of myeloid differentiation	literature co-occurrence (1.0)
MEIS2	negative regulation of myeloid differentiation	literature co-occurrence (0.31)
HOXA5	positive regulation of myeloid differentiation	literature co-occurrence (1.0)
HOXB4	hematopoietic stem cell differentiation	literature co-occurrence (0.64)
HOXB9	negative regulation of myeloid differentiation	literature co-occurrence (0.11)

The HOXA9/MEIS1 genes above were one slice of what that single question returned. The same query also surfaced a second, broader group: genes with deep, multi-evidence links to the same arrest program that do not sit on the HOXA9/MEIS1 axis itself.

Gene	GO role	Notable AML evidence
CEBPA	myeloid cell differentiation	curated causative, clinically validated, driver mutation (all 1.0)
ZBTB16 (PLZF)	neg. + general myeloid differentiation	somatic mutation (0.75), literature (1.0)
KLF4	stem cell population maintenance	somatic mutation, drug target, literature — all high
CTNNB1 (β-catenin)	stem cell population maintenance	somatic mutation, drug target, literature — Wnt/stemness pathway
KAT6A (MOZ)	myeloid cell differentiation	somatic mutation (0.75), literature (1.0) — MOZ fusions block differentiation
TET1/TET2	myeloid & hematopoietic differentiation regulation	TET2 somatic mutation, driver mutation, burden (all 1.0)

The same query also returned genes with comparatively weak, secondary evidence for the same association - NCOA6 and IFI16 among them flagged as such rather than silently folded in at equal weight.

KEY FINDING

HOXA9 and MEIS1 co-expression is the textbook leukemogenic signature, classically driven by KMT2A rearrangements and NUP98 fusions, and the graph's own annotations line up with that biology exactly, tagging both genes independently as negative regulators of myeloid differentiation.

Resolving to a genetic subtype: NPM1-mutant AML

AML is not one disease in the graph. Narrowing the query to NPM1-mutant AML specifically returned a distinct disease node -not the general AML term, connected to NPM1, MEN1, and HOXA9. The direct disease-to-gene edges, though, told only part of the story:

Edge	Edge type	Score
AML with mutated NPM1 → NPM1	has_somatic_mutation_in	**1.0**
AML (general) → MEN1	has_dependency_on	**0.42**
AML → MEN1	has_drug_target	**0.70**
AML → MEN1	has_somatic_mutation_in	**0.75**
AML → HOXA9	has_somatic_mutation_in	**0.75**

HOXA9, MEIS1, and MEN1 don't connect directly to the NPM1-mutant AML node - only NPM1 does.

WHY THIS MATTERS

That's not a gap in the data. It's how the biology actually works: the NPM1 mutation doesn't touch HOXA9, MEIS1, or MEN1 directly. Instead, it produces a mislocalized, cytoplasmic form of NPM1 that keeps the menin–KMT2A complex active, and that complex is what keeps HOXA9 and MEIS1 switched on. Reaching that mechanism meant querying the gene-gene interaction layer next.

The gene-gene bridge

Edge	Type	Score
MEN1 → HOXA9	methylation	—
MEN1 → KMT2A	association	**0.64**
NPM1 → HOXA9	interaction	**0.454**
NPM1 → MEIS1	interaction	**0.441**

Each gene's own GO annotations turn this from a list of edges into an actual mechanism. HOXA9 is tagged for negative regulation of myeloid differentiation. MEIS1 is not - its annotations cover cell-cycle and transcription control, which fits it acting upstream of HOXA9 rather than on differentiation directly. KMT2A is tagged for definitive hemopoiesis and epigenetic regulation.

NPM1's own annotations - chromatin remodeling, ribosome biogenesis, general proliferation control describe its normal job, and none of them mention differentiation arrest, which fits NPM1 acting indirectly through menin rather than driving the arrest itself.

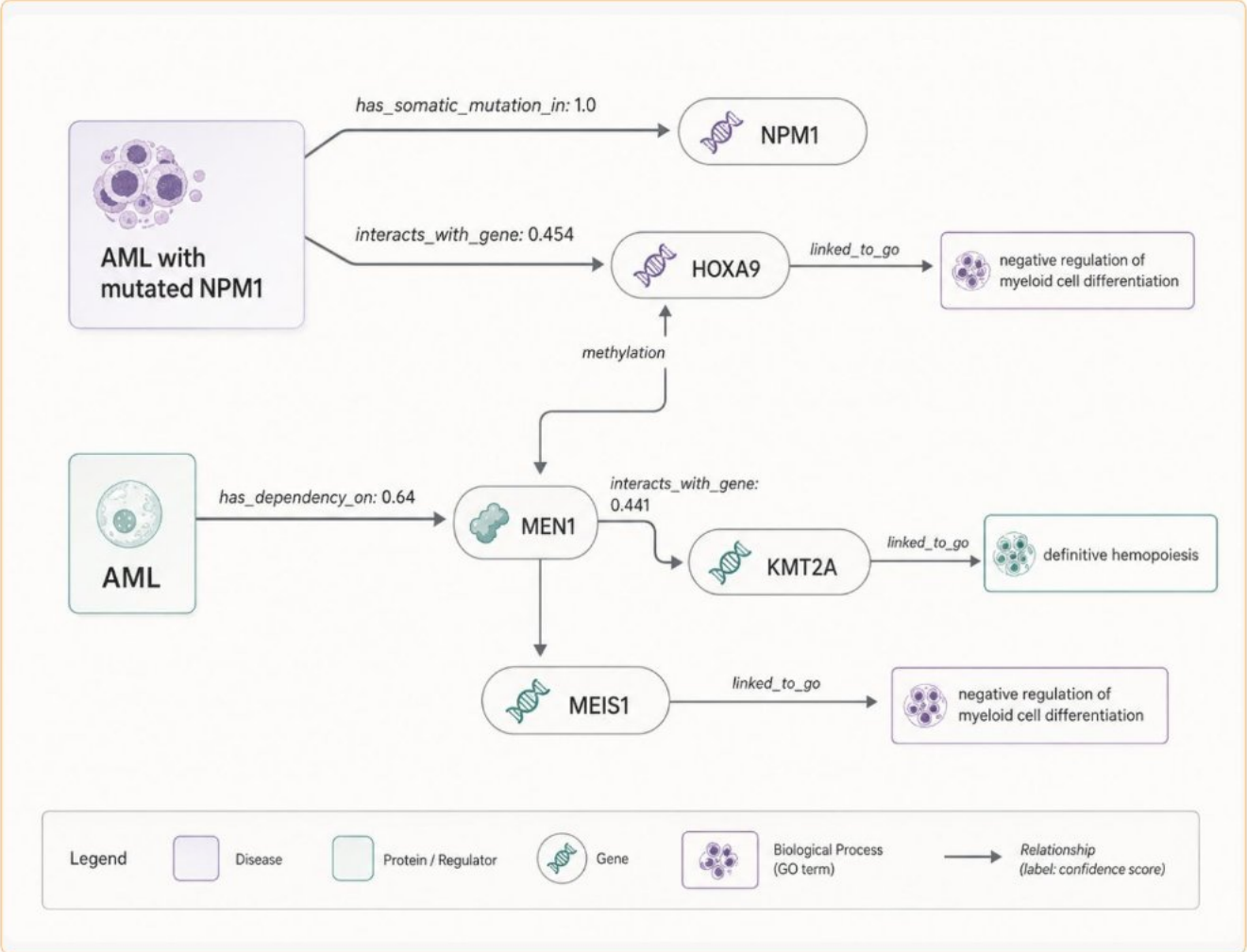


Figure 4. The assembled path: AML with mutated NPM1 reaches HOXA9 and MEIS1 through two routes, direct interaction and the MEN1–KMT2A bridge, both converging on negative regulation of myeloid cell differentiation.

KEY FINDING

This is the well-established AML mechanism, recovered from graph structure rather than restated from memory: the mislocalized NPM1 protein sustains the menin–KMT2A complex, which keeps HOXA9 and MEIS1 switched on. That is the molecular basis for the differentiation block that menin inhibitors are designed to disrupt.

Druggability and tractability

With the mechanism established, the next question was simple: which points on this path are actually druggable, and how far along?

Gene	Tractability	Drug (phase/approval)	Mechanism
MEN1	phase_1_clinical, high_quality_ligand, small_molecule_binder	Revumenib - Approved 2024	Menin–KMT2A interaction inhibitor
KMT2A	phase_1_clinical, druggable_family, high_quality_ligand	Revumenib (same drug, shared target) — Approved 2024	Disrupts the menin–KMT2A association

MEN1 and KMT2A share the same drug because they are the same drug target - revumenib works by disrupting the physical interaction between them, not by acting on either gene in isolation.

Gene	Drug	Phase	Mechanism
FLT3 (co-mutated in ~50% of NPM1-mutant AML)	Midostaurin, Quizartinib	Approved (2017, 2023)	FLT3 tyrosine kinase inhibition
IDH1	Olutasidenib	Approved 2022	Mutant IDH1 inhibitor
IDH2	Enasidenib	Approved 2017	Mutant IDH2 inhibitor
DNMT3A	Azacitidine	Approved 2004	Hypomethylating agent (epigenetic)

None of these four sit on the direct NPM1→HOXA9/MEIS1 path. They are, instead, the dominant co-mutation partners of NPM1 in real AML patients, which is exactly why combination strategies pairing a menin inhibitor with one of these are clinically relevant rather than speculative.

Gene	Drug	Phase	Note
CDK6	Alvocidib	Phase 3	Pan-CDK inhibitor, not selective
BRD4	Birabresib	Phase 1/2	BET inhibitor — disrupts HOXA9/MEIS1 transcription indirectly via chromatin-reader inhibition

CDK6 and BRD4 sit further out still - earlier-stage, less selective, or working on the axis indirectly rather than on MEN1 or KMT2A themselves.

KEY FINDING

The graph connected the NPM1c–HOX/MEIS1 axis to the MEN1–KMT2A axis through the MLL1 chromatin-regulatory complex, revealing a shared mechanism that sustains differentiation arrest in NPM1-mutant AML.

Revumenib, an approved menin inhibitor, sits directly on this axis and is already active in NPM1-mutant and KMT2A-rearranged AML. The same query also surfaced two less finished stories worth keeping in view. DOT1L, in the same regulatory neighborhood, has only an investigational compound- pinometostat, still in AML trials, not yet approved which makes it more a next-generation candidate than a current one.

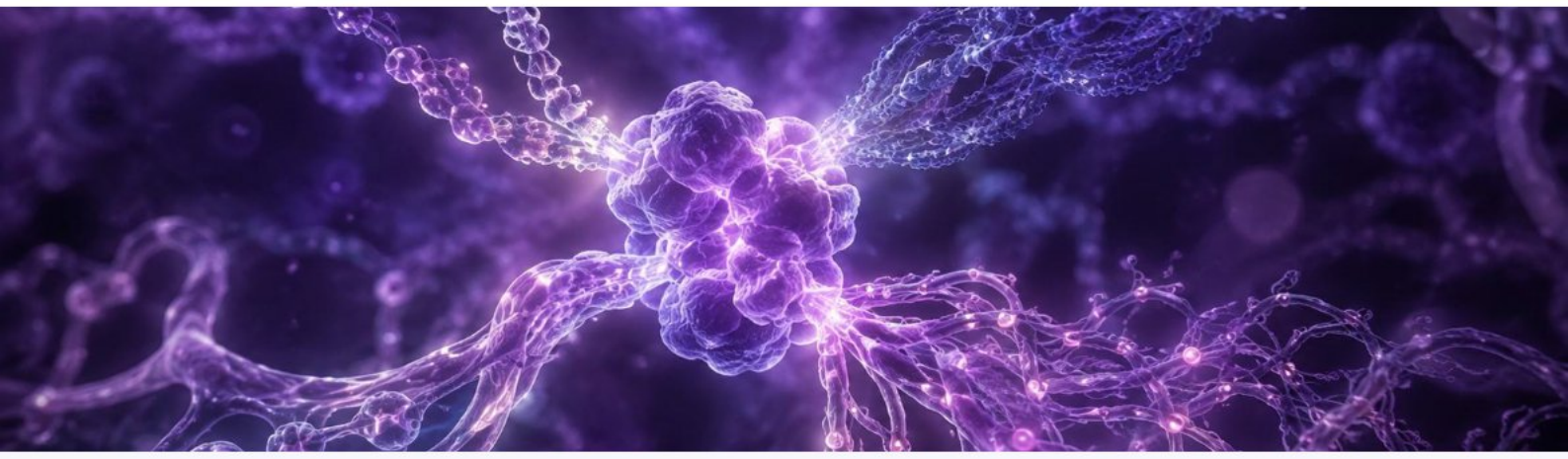
MEIS1, the transcription factor sitting at the center of this axis, has no compound in the graph at all, which tracks with how hard transcription factors are to drug directly rather than any gap in the evidence itself.

Gene / complex	Compound	Status
MEN1 / KMT2A	Revumenib	Approved menin inhibitor
DOT1L	Pinometostat	Investigational — evaluated in AML trials
MEIS1	—	No direct compound; transcription factor

Physical complex and pathway convergence

MEN1 and KMT2A have no direct gene-gene interaction edge in the graph. What the graph does encode, directly, is that they are subunits of the same physical protein assemblies:

Complex	Gene
MLL1 complex	**MEN1 + KMT2A (co-subunits)**
MLL–HCF complex	**MEN1 + KMT2A (co-subunits)**
ASH2L–DPY30–KMT2A–RBBP5–WDR5 complex	**KMT2A**
MLL1–WDR5 complex	**KMT2A**
Menin-associated histone methyltransferase complex	**MEN1**



That shared complex membership is the molecular basis for menin-inhibitor drugs disrupting the interaction in the first place. A second layer, Reactome pathway membership, reinforces the same convergence from a different angle:

Pathway	Gene
Formation of WDR5-containing histone-modifying complexes	**MEN1 + KMT2A (shared)**
PKMTs methylate histone lysines	**KMT2A**
Transcriptional regulation of granulopoiesis	**KMT2A**
RUNX1 regulates transcription of genes involved in HSC differentiation	**KMT2A**
RUNX1 regulates genes involved in megakaryocyte differentiation	**KMT2A**

These pathway edges place KMT2A directly inside hematopoietic stem-cell and myeloid differentiation regulatory programs.

KEY FINDING

Both genes converge on the same histone-modifying complex-formation pathway - the epigenetic mechanism underneath the physical one. Two separate evidence layers, physical complex and shared pathway, point at exactly the same convergence from different angles.

Functional validation through CRISPR dependency

A fourth, independent layer - functional perturbation data from CRISPR knockout screens — tested whether the axis holds up experimentally, not just structurally:

Gene	Cell line	Gene effect (more negative = more essential)
MEN1	MOLM-13 (KMT2A-rearranged AML)	**−2.18**
MEN1	HT	**−2.21**
MEN1	KMS-20	**−2.14**
KMT2A	MOLM-13	**−0.90**
KMT2A	NOMO-1	**−0.68**
KMT2A	MOLM-14	**−0.61**
KMT2A	THP-1	**−0.45**

MOLM-13 and MOLM-14 are the canonical KMT2A-rearranged AML cell lines, and MEN1 shows a substantially stronger dependency signal than KMT2A itself in these lines.

A fifth layer, disease-association breadth, adds one more piece of context: KMT2A's evidence spans a strikingly wide range of leukemia subtypes - AML, ALL, CML, JMML, and acute leukemia of ambiguous lineage.

Put together, six independent evidence layers - disease association, physical complex membership, shared Reactome pathway, gene ontology, direct gene-gene interaction, and functional CRISPR dependency converge on the same axis from six different directions, none of which alone would have been conclusive.

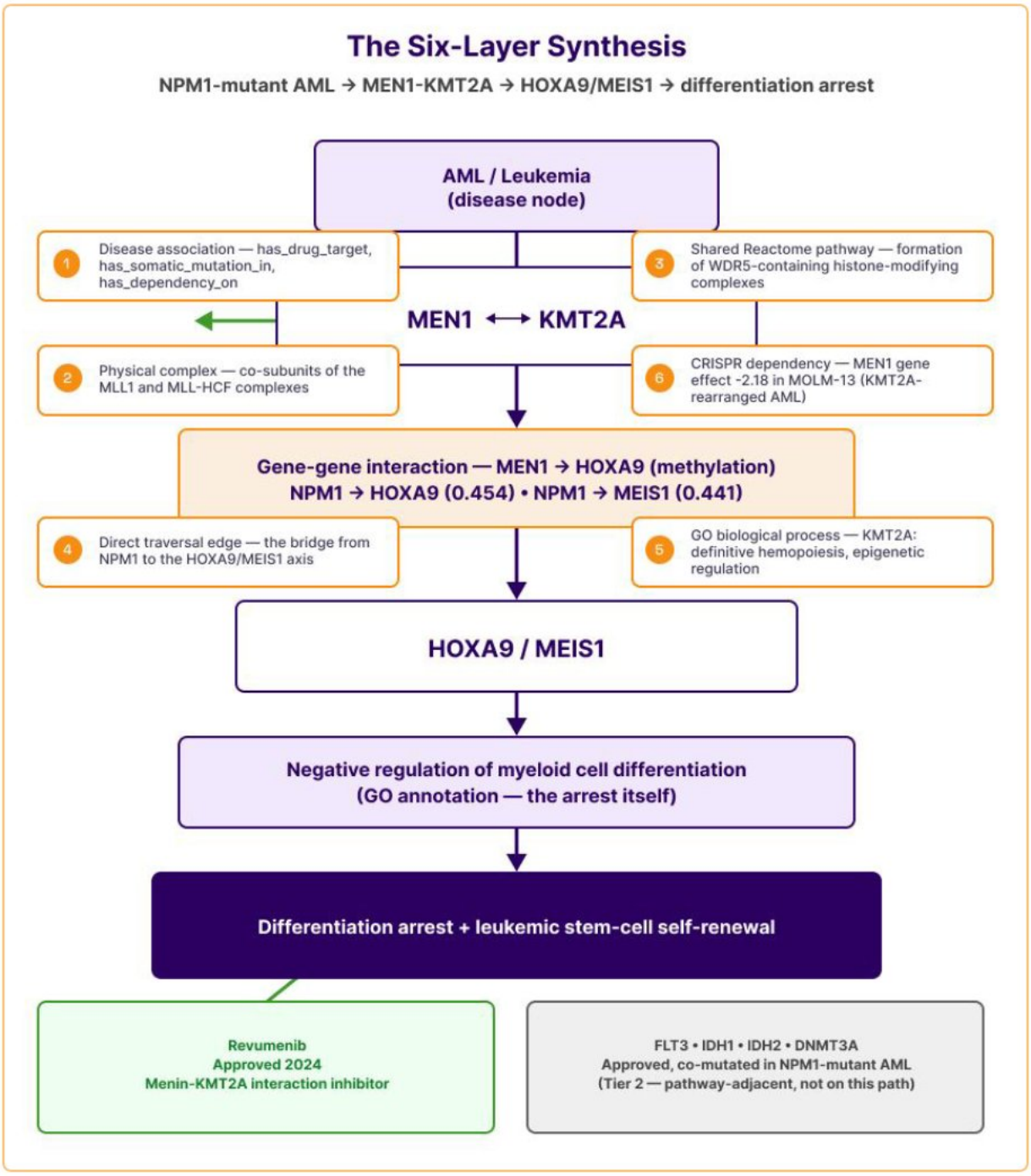


Figure 5. Six independent evidence layers converging on the MEN1-KMT2A axis and the HOXA9/MEIS1 differentiation-arrest

These facts are not new to AML biology individually. What changes is that a single natural-language session, working layer by layer through a graph which reconstructed the connection between them - in place of the six independent screens, cohorts, and publications that assembling this picture would ordinarily require by hand.

Key Takeaways

01

Mechanistically typed and connected evidence increases confidence in target prioritization by removing the manual step required to confirm that a co-occurrence reflects a real biological association.

02

The layered graph architecture and scoring methodology used for AML here - evidence-type ranking, tissue specificity, gene ontology, tractability, protein-complex membership, and CRISPR dependency apply without modification to other disease areas, data-sparse ones included.

03

A single disease-specific knowledge graph supports biomarker identification, companion diagnostic development, and indication expansion from the same underlying evidence base.

04

Most importantly: none of this required a fresh literature synthesis. Work that would ordinarily take a computational biology team days - pulling records from separate databases and reconciling them against primary literature manually, happened in a single natural-language session, with every claim traceable to source.

Future Roadmap

Temporal modeling of disease progression

The current graph represents biological relationships as static or directional edges. Extending this to represent disease-state transitions over time is a defined next step, particularly relevant to diseases such as AML, where resistance and relapse involve changes in cellular state rather than fixed genetic lesions.

Multi-step, agent-driven analysis

Today MCP answers one question at a time. Each query is handled on its own: it does not plan a multi-step investigation, chain several questions together, or decide what to ask next based on what the last answer returned. In this case study, that sequencing was the scientist's job. Building agents that can carry out that sequencing themselves, running functional enrichment analysis or cross-subtype comparison across a chain of dependent questions, is planned.

Putting query design in domain scientists' hands

The value of a conversational interface to the knowledge graph depends on how precisely a domain scientist, rather than a computational specialist, like how a question should be interpreted, which evidence layers to consult, which tools to call. Those instructions take different forms across assistants, skill files in some, comparable configuration formats in others. This is an adoption and training consideration as much as a technical one.

Conclusion

NPM1, HOXA9, MEIS1, KMT2A, MEN1, and revumenib were all independently known before this exercise began. What did not exist anywhere was the connected version: a single traceable path from a disease name to an approved, genetically matched drug, assembled from evidence that ordinarily sits across five or six unrelated sources and cites none of them together.

That is the output worth measuring. Not new biology, but a defensible target thesis produced in a fraction of the time a computational biology team would spend reconciling databases and primary literature by hand, with every claim traceable to source when the thesis reaches review. The scientific value is a mechanism that holds together across six independent evidence layers. The commercial value is that the work took under half an hour rather than several days, and produced a result someone can actually defend.

This framework extends to any disease where the mechanism is real but scattered - surfacing repurposing opportunities through shared pathways and connecting known biology across indications to find new targets, even where data is genuinely scarce.

References

1. Xi Y, Zhang T, Sun W, Liang R, Ganesh S, Chen H. GOLM1 and FAM49B: potential biomarkers in HNSCC based on bioinformatics and immunohistochemical analysis. Int J Mol Sci. 2022;23(23):15433. PMID: 36499755.
2. Nelson MR, Tipney H, Painter JL, et al. The support of human genetic evidence for approved drug indications. Nat Genet. 2015;47(8):856–860.
3. Ochoa D, Hercules A, Carmona M, et al. The next-generation Open Targets Platform: reimagined, redesigned, rebuilt. Nucleic Acids Res. 2023;51(D1):D1353–D1359.
4. Tsherniak A, Vazquez F, Montgomery PG, et al. Defining a cancer dependency map. Cell. 2017;170(3):564–576. (DepMap)
5. Gillespie M, Jassal B, Stephan R, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022;50(D1):D687–D692.
6. U.S. Food and Drug Administration. FDA approves revumenib for relapsed or refractory acute myeloid leukemia with a susceptible NPM1 mutation. October 24, 2025.

Glossary

AML - Acute Myeloid Leukemia. An aggressive hematological malignancy characterized by impaired differentiation of hematopoietic stem and progenitor cells.

NPM1 / NPM1c - NPM1 is the most commonly mutated gene in adult AML; NPM1c refers to the resulting mislocalized, cytoplasmic mutant protein. NPM1-mutant AML is a distinct disease-subtype node in the graph.

KMT2A - A gene encoding a histone methyltransferase; a recurrent fusion/rearrangement partner across multiple leukemia subtypes, and a co-subunit of the MLL1 complex with MEN1.

MEN1 - Encodes menin, a chromatin-associated protein and validated druggable target within the MLL1 complex in AML.

Menin inhibitor - A drug class that blocks the menin–KMT2A interaction; revumenib is an approved example, active in genetically defined AML subsets including NPM1-mutant disease.

MLL1 / MLL–HCF complex - Chromatin-regulatory protein complexes in which MEN1 and KMT2A are co-subunits, despite having no direct gene-gene interaction edge between them.

HOXA9 / MEIS1 - Homeobox and cofactor genes whose co-expression is the classic leukemogenic signature; both are independently annotated as negative regulators of myeloid differentiation.

Tractability annotation - A graph property (e.g., phase_1_clinical, druggable_family, high_quality_ligand) that records how far a target already is along the path to a compound, tracked independently of disease-association strength.

Gene-effect score - A CRISPR knockout-screen readout of how essential a gene is to a specific cell line's viability; more negative values indicate stronger functional dependency.

Reactome pathway - A curated biological-pathway annotation; shared Reactome membership between two genes is one of several ways the graph can connect them without a direct interaction edge.

Directed subgraph - The output format MCP returns for multi-hop questions — a traceable path through several nodes and typed edges, rather than a single row or score.

Polly KG - The underlying knowledge graph: roughly 12.44 million nodes across 13 node types and 40.39 million edges across 183 typed, directional edge types, layered with disease-specific and proprietary evidence.

Polly MCP - A Model Context Protocol connector that lets an AI assistant query Polly KG using natural language, directly from Claude, GPT, Gemini, or an enterprise model.

Skill file - A configuration that defines how an AI assistant should interpret a question, which steps to follow, and which tools to use against a given knowledge graph.

About Elucidata

Elucidata co-builds AI solutions with real biomedical data. Unstructured, siloed data (molecular, clinical, imaging, EHR/EMR, and 25+ other modalities) is transformed through the Polly platform into tailored AI solutions for novel target discovery, pre-IND reporting, biomarker identification, tumor board support, and portfolio intelligence.

40	10	7	4
programs past IND	therapeutic areas	drug modalities	approved drugs

Deployed on a secure cloud (HIPAA, GDPR, SOC 2 compliant)

Have questions or want to discuss the ideas shared in this whitepaper?

We'd be happy to hear from you.

For any queries, clarifications, or to explore how these approaches could apply to your work, please reach out to us at info@elucidata.io

Book a Demo