Predicting Novel Crosstalks in Oncology Using Knowledge Graphs

Our approach to uncovering Disease biology for Data-sparse Cancers, using evidence-rich knowledge graphs

WHITEPAPER

Contents

Executive Summary	03
Introduction: Why Target Discovery Remains Difficult	03
The Solution: Modeling Disease as a Connected System	05
Case Study: Predicting a Druggable Marker for Neuroendocrine Prostate Cancer (NEPC)	06
Building the NEPC causal knowledge graph	08
The Open Targets Challenge: Gaps in Disease-Specific Biology	09
Target Prioritization: Co-Essentiality and Link Prediction	10
Link prediction revealed an epigenetic crosstalk	11
The Drug Development Landscape in the PRC2 cluster	12
Biological and Clinical Validation	13
EHMT2: External Literature Convergence	13
Future Directions	14
Conclusion	15
References	16
Glossary	16

Executive Summary

The promise of knowledge graphs lies in their ability to cover a large set of biological relationships - across Genes, Disease, Pathways and Proteins.

Commercial knowledge graphs are often evaluated on the basis of the breadth of coverage that they have. Public literature is an important source of information for these knowledge graphs - Open Targets incorporates co-occurrence relationships from over ~10M research papers.

In this white paper, we talk about how we've built an evidence-rich knowledge graph that helped identify a potential druggable marker for Neuroendocrine Prostate Cancer. This is a subtype of prostate cancer that afflicts around ~20,000 people in the US every year with median survival rate of 7 months, and emerges from neuroplasticity-driven adaptation to the standard course of treatment to metastatic castrate-resistant prostate cancer (CRPC) patients.

Our analysis predicted a novel epigenetic crosstalk between LCOR and EHMT2 within the PRC2 chromatin-regulatory axis. This crosstalk, invisible to standard databases and absent from the initial target list, revealed EHMT2 (G9a) as a druggable chromatin regulator with active pre-clinical drug chemistry. This makes it a more actionable target than MYCN itself. The finding was later supported by functional evidence from related cancers and presented at AACR in April 2026.

This methodology is essentially an approach to fleshing out regulatory cross-talk in scenarios where public data is fragmented and for rare diseases with very sparse evidence available.

Introduction: Why Target Discovery Remains Difficult

Targets supported by strong mechanistic genetic evidence are more than twice as likely to succeed in clinical trials.

Most workflows rely on experimental data repositories (e.g., TCGA, DepMap, cBioPortal) and start with foundational platforms for aggregating literature (e.g., Open Targets). However, these experimental repositories are often fragmented and lack data for underrepresented cancers. And literature platforms cannot distinguish between a true, driving causal biological mechanism and a purely incidental mention in a research. This gap is where target prioritization programs lose their clinical advantage.

Three structural challenges that make Target Prioritization hard:

1. Co-occurrence is not Biological Evidence

Open Targets platform integrates approximately 30 biomedical entity types with continuous monthly updates to generate roughly 21 million evidence pairs from over 10 million abstracts. However, the core limitation is that its evidence layer is flat and based on co-occurrences - instances where a gene and a disease are simply mentioned in the same text.

A sentence explicitly reporting "no evidence for association" is captured identically to a sentence reporting a "strong association" because both contain the entity names. The system does not extract the mechanistic trail connecting entities through causal pathways.

Furthermore, the system is primarily scoped for gene-disease pairs, restricting the multi-hop reasoning needed for gene-pathway or cross-entity connections. It also misses critical data from approximately 8 million full-text papers whose methods sections, supplementary data, and mechanistic discussions are not captured in the abstract-based evidence layer.

2. The Prioritization Funnel Lacks a Systematic Framework

Every drug discovery program starts with the same search space: identifying targets from approximately 20,000 protein-coding genes and narrowing them down to 2 viable candidates.

Without a structured, evidence-based ranking framework, it is difficult to shortlist targets based on evidence.

This problem is heavily compounded by data scarcity in rare or underrepresented cancer subtypes. Models trained on data-rich cancers produce outputs that appear robust but fail to transfer reliably to these indications, and proprietary experimental data cannot be easily integrated into most off-the-shelf platforms.

3. Current State-Of-Art KGs Lack Mechanistic Insights

To navigate these bottlenecks, many teams rely on generic off-the-shelf knowledge graphs The key point commercial off-the-shelf knowledge graphs is the range of relationships available. However, the lack of a focus on a specific disease biology, in addition to an architecture that does not allow proprietary data and analysis to be included, make these knowledge graphs at best a science experiment, but not the foundations of a serious research program.

Ultimately, these generic platforms answer broad biological queries rather than the precise therapeutic questions required to confidently advance a unique drug discovery program.

The Solution: Modeling Disease as a Connected System

It is our strong belief that biology must be modeled around a specific therapeutic question rather than general data aggregation. In data-scarce environments, disease-specific knowledge graphs can enable powerful inference by connecting biological signals through shared causal mechanisms of other diseases or subtypes and pathways to find repurposing opportunities and novel crosstalks.

Even when direct information for a disease is missing, core mechanisms such as genetic dependencies, epigenetic machinery, and functional CRISPR screens remain conserved across indications.

Capturing these conserved mechanisms and transferring these signals is necessary but not sufficient. The critical differentiator is whether those mechanisms are weighted, scored, and connected in the context of a specific disease, which requires disease-specific data integration and a custom scoring framework that generic tools do not provide.

Our Approach: Polly Knowledge Graph (Polly KG)

Polly KG operationalizes this approach by integrating over 20 public sources, 8 million full-text papers, and proprietary data into a unified, ontology-mapped network. Every relationship is typed and directional: upregulation is distinguished from downregulation, and protein-protein interactions are distinguished from regulatory relationships. Multi-source confidence scoring is applied to every edge.

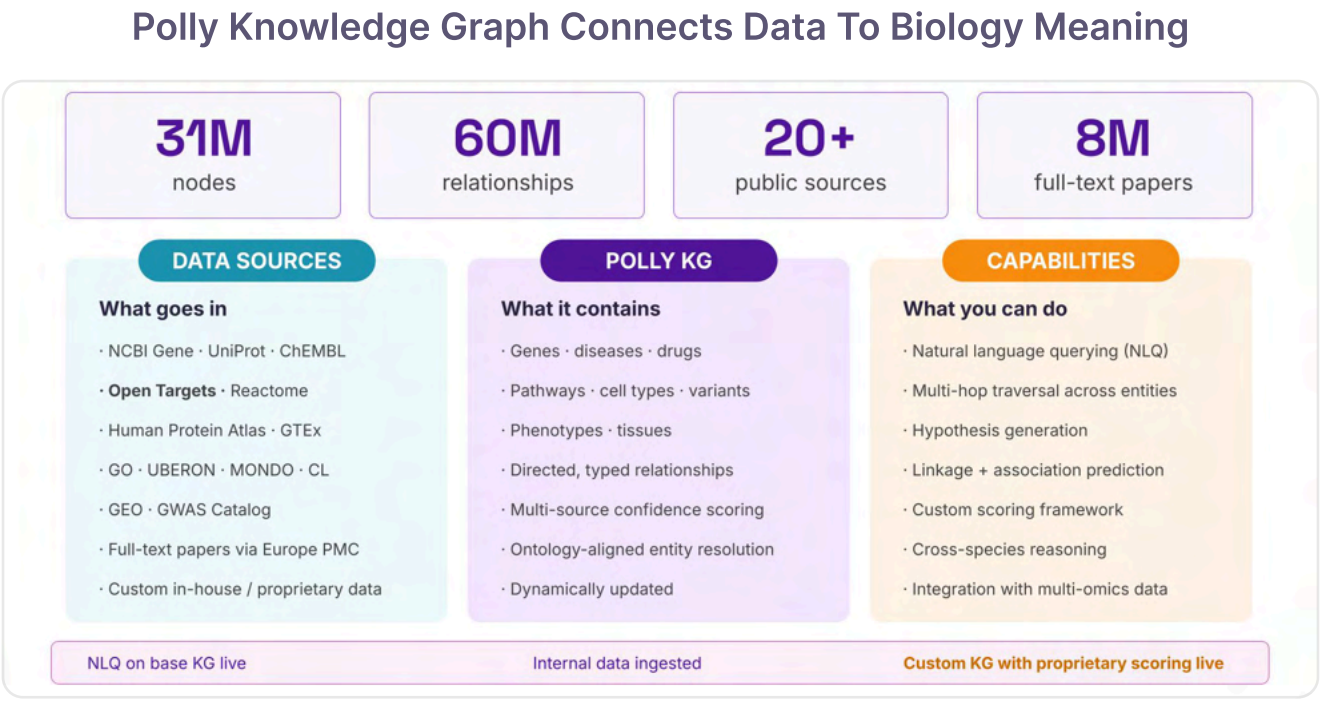


Figure 1. Polly KG architecture: data sources (what goes in), graph contents (what it contains), and capabilities (what can be done). The system supports natural language querying from Day 1.

Moving beyond data retrieval, Polly KG deploys two advanced capabilities to drive target discovery:

1. Custom Scoring:

Applies flexible, disease-specific weights to biological dimensions like tissue context, pathway involvement, and molecular tractability, ranking targets by their precise contextual relevance.

2. Edge Prediction:

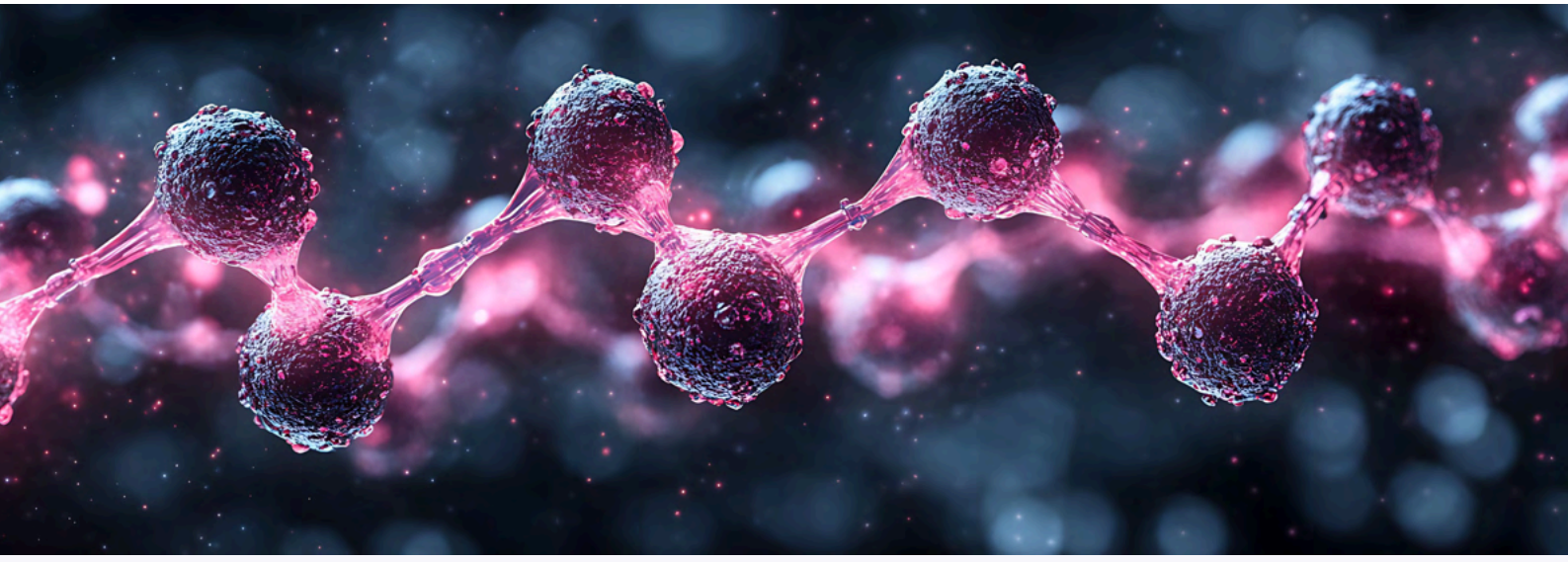
Combines graph neural network embeddings with these scoring frameworks to infer previously unobserved gene-disease or cross-entity connections, accelerating hypothesis generation beyond published literature.

By deploying these capabilities, this document focuses on three high-value applications: predicting novel molecular interactions, identifying drivers of tumor progression and assessing whether approved drugs in one cancer context may modulate mutation-driven disease states in another to find drug repurposing and therapeutic expansion opportunities.

Case Study: Prioritizing EHMT2 as a Druggable Marker for NEPC

The practical execution of this framework is demonstrated through our case study presented at AACR for Neuroendocrine Prostate Cancer (NEPC), an aggressive disease state where the lack of public data makes it very hard to initiate and prioritize a research programme.

The key takeaway is that the target prioritized in this case study does not emerge from querying usual databases. It resulted from integrating multi-modal data to model a specific biological transition, leading to a biologically validated, druggable marker with real published evidence from adjacent cancers that this intervention actually works leading to a significantly shortened path from nomination to drug development.



The Challenge :

1. Treatment-induced lineage plasticity in NEPC

Prostate cancer cells normally depend on the androgen receptor (AR). When targeted with therapies like enzalutamide, most cells die. A subset survives by abandoning the prostate lineage entirely and acquiring neuroendocrine characteristics. The drug target disappears because the disease itself has transformed and, this leads to the loss of conventional targets.

Cell identity in this context is not encoded in DNA sequence but in chromatin state, which genes are accessible and which are silenced. The master regulator of the neuroendocrine transition is MYCN, which, in the absence of AR signaling, recruits the PRC2 chromatin-modifying complex to silence the prostate transcriptional program and activate the neuroendocrine program. MYCN, however, is a transcription factor with no druggable binding pocket. No clinical MYCN inhibitor has reached approval.

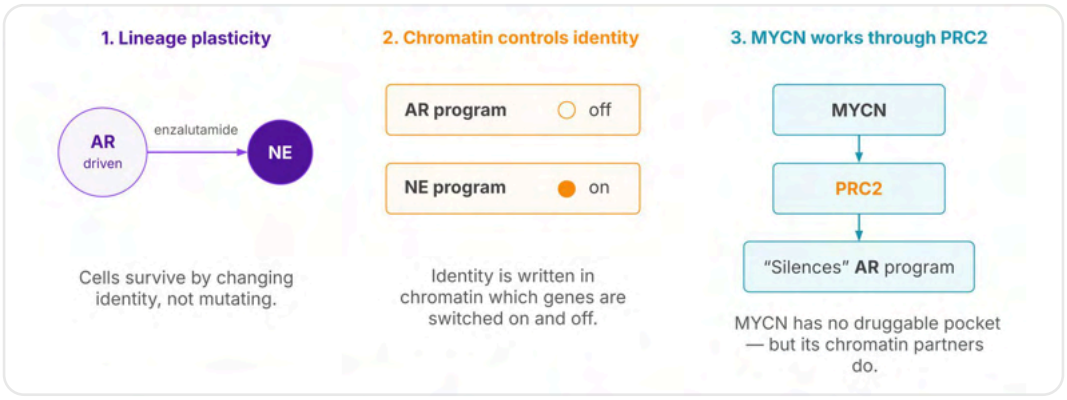


Figure 2. NEPC emerges under treatment pressure through lineage plasticity. Cells survive by changing identity rather than acquiring new mutations. The chromatin-level identity switch is driven by MYCN through the PRC2 complex.

Strategy

If MYCN cannot be targeted directly, can the identity switch be modeled computationally to identify druggable chromatin partners within the same regulatory axis?

2. Structural Data Scarcity:

A primary reason early-stage discovery is so difficult is the severe lack of available NEPC studies. The disease is systematically underrepresented across foundational oncology repositories. TCGA excludes neuroendocrine histology by design, COSMIC provides only histology-tagged mutation counts without matched expression data, and cBioPortal offers only a single dataset with no treatment-stage cohorts.

Database	NEPC Data	Limitation
TCGA	No NEPC data	Excludes neuroendocrine and small-cell histology by design
COSMIC	2 publications	Histology-tagged mutation counts; no matched expression data
GEO	Available	Data is neither findable nor usable without significant curation
cBioPortal	1 dataset	Treatment-stage cohorts not available

Building the NEPC causal knowledge graph

The first step in constructing the NEPC-specific graph was identifying the datasets missing from standard databases.

We deployed Polly Scout, an AI-driven search layer across more than one million public records and queried for prostate cancer subtypes with treatment-stage stratification. The system rapidly filtered 300 initial datasets down to 10 highly relevant temporal cohorts, specifically isolating the exact biological windows where the lineage transition occurs in few hours cutting down weeks of manual browsing and metadata review.

These datasets were then integrated into Polly KG through a three-layered modeling framework to build a disease-specific causal network for evaluating targets-

Layer 1: Foundational biological knowledge (Base KG)-Pathway, disease, ontology, and molecular relationship context drawn from NCBI, UniProt, PubMed, Reactome, Human Protein Atlas, GTEx, and ChEMBL. This layer is available from Day 1. Most off-the-shelf knowledge graphs provide only this layer.

Layer 2: NEPC contextual biology- The disease-specific datasets identified through Polly Scout (treatment-stage cohorts from GEO, TCGA, and cBioPortal, along with DepMap essentiality data, LINCS perturbation profiles, and lineage transition datasets) were ingested through automated ETL pipelines, mapped to a custom data model, and quality-checked before integration. This layer captures the co-essentiality signals and perturbation context specific to the NEPC transition.

Layer 3: Mechanistic prioritization-Deploys co-essentiality mapping and link prediction algorithms directly on this multi-layered architecture to enable custom mechanistically grounded target ranking. The graph reasons beyond observed evidence to infer novel crosstalks and answer all the queries about the specific biological hypothesis in Natural Language.

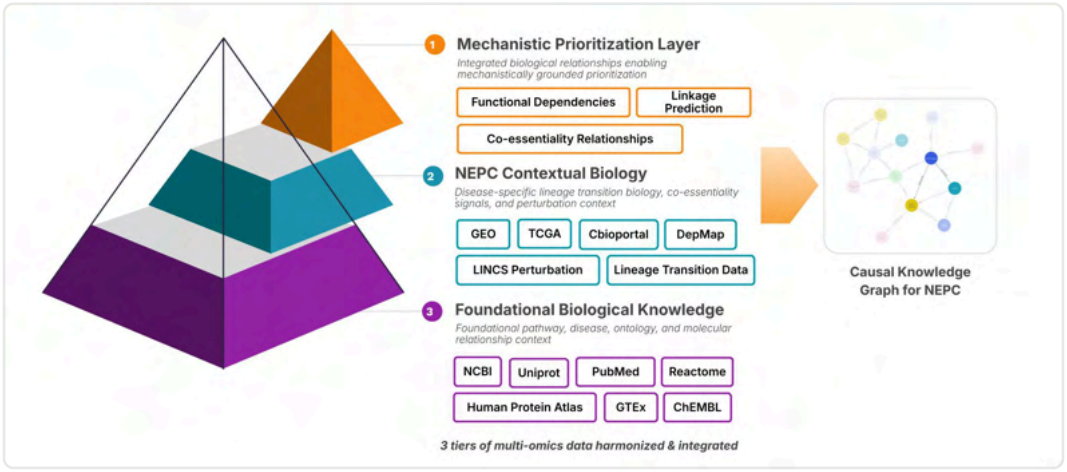


Figure 3. Three-layered NEPC causal knowledge graph built from harmonized biological evidence. Foundational knowledge, NEPC contextual biology, and a mechanistic prioritization layer are integrated from distinct data source tiers.

Because the graph enforces directional edges throughout, multi-hop reasoning is possible: a single query can traverse from a clinical observation to a specific molecular mechanism to a relevant patient cohort.

The Open Targets Challenge: Gaps in Disease-Specific Biology

Querying the foundational data layer demonstrated its value as a baseline filter, identifying approximately 35,000 prostate cancer evidences, roughly 400 of which mentioned NEPC. However, when evaluated against strict druggability and chromatin-axis criteria, only one candidate remained: EZH2. Because EZH2 is already targeted by an approved drug (Tazemetostat) and tumors rapidly diversify resistance mechanisms beyond single interventions, this target offered no novel path forward.

Similarly, querying Open Targets directly returned only weak, literature-based associations for the PRC2 cluster in generic prostate cancer, completely missing the NEPC-specific context. This is a structural limitation, not a flaw in the database: NEPC-specific mechanistic crosstalks are simply not published as discrete gene-disease pairs. The relevant evidence is fragmented across essentiality screens, expression datasets, and disconnected temporal cohorts. These are sources that flat databases and off-the-shelf tools cannot integrate.

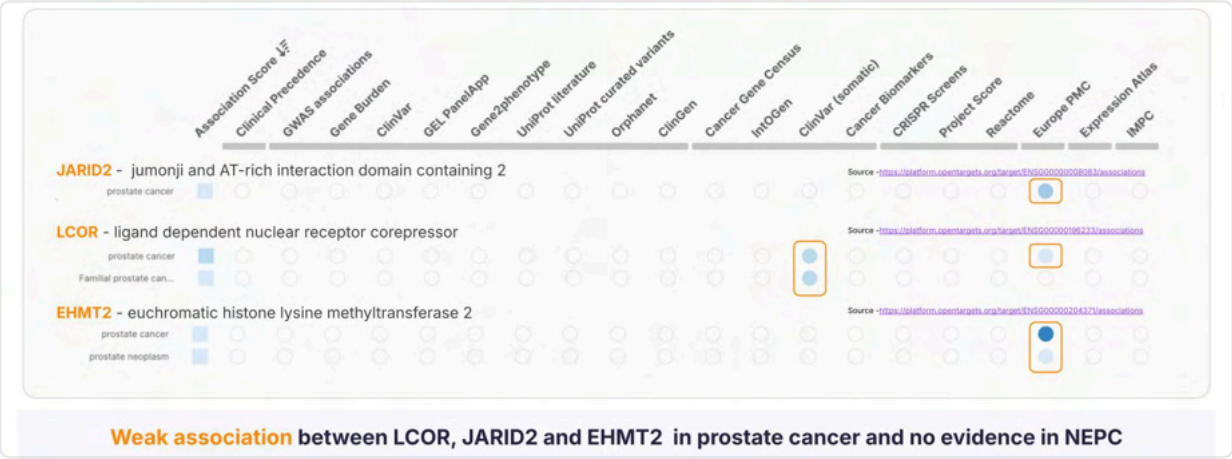


Figure 6.Open Targets evidence for JARID2, LCOR, and EHMT2 in prostate cancer. Weak association scores, predominantly from literature co-occurrence. No NEPC-specific evidence across any data source.

Target Prioritization: Co-Essentiality and Link Prediction

To move beyond the established targets that standard databases provide, the Polly Knowledge Graph (Polly KG) anchored its analysis on MYCN, a master transcription factor that drives the tumor's transition to a neuroendocrine state but lacks a viable binding pocket for direct drug intervention. Because MYCN itself is undruggable, the graph mapped its surrounding regulatory network.

By deploying its mechanistic layer across the NEPC-specific cohorts, the Knowledge Graph successfully mapped co-essentiality and surfaced two critical PRC2 complex components demonstrating a clear regulatory hand-off:

Gene	Co-essentiality	Significance
JARID2	r = 0.93 with MYCN	Tumor suppressor; downregulated in post- and mid-treatment cohorts. Part of PRC2 complex.
LCOR	r = 0.90 with MYCN	Co-repressor; upregulated during late treatment. Part of PRC2 complex. Accelerates forward differentiation.

Link prediction revealed an epigenetic crosstalk

Operating on this structural evidence, the Knowledge Graph inferred a previously unreported edge: a predicted epigenetic crosstalk between LCOR and EHMT2. The Knowledge Graph found this connection to be significant for four reasons:

- **Structural Convergence:** EHMT2 did not appear in the initial target list; it surfaced because the graph recognized that chromatin organization, the PRC2 complex, and MYCN co-essentiality all structurally converged upon it.
- **Chromatin Machinery:** EHMT2 encodes a protein within the same chromatin-regulatory machinery as PRC2, supported by domain-level evidence natively integrated into the graph.
- **Algorithmic Discovery:** It was completely absent from the initial co-essentiality scan, requiring the graph's link prediction capabilities to be discovered.
- **Temporal Alignment:** Both EHMT2 and LCOR were upregulated in the same late-treatment cohorts, indicating a functional crosstalk contributing to resistance.

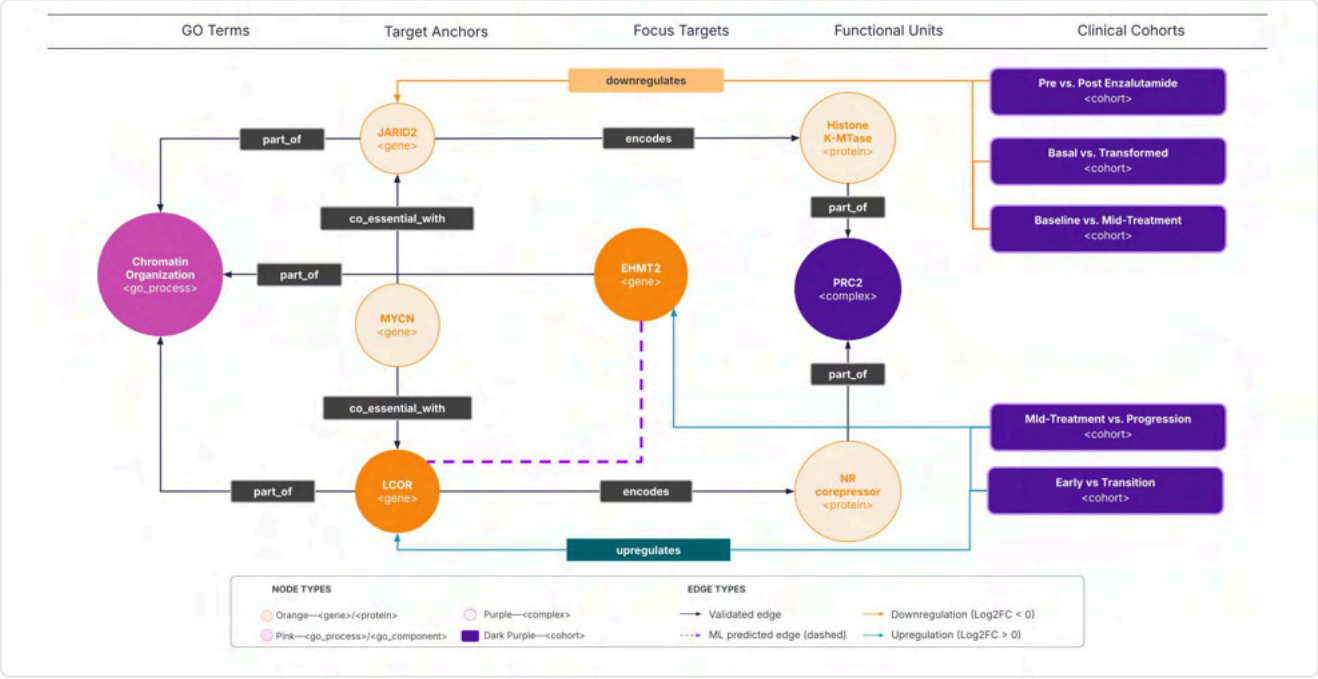


Figure 4.Link prediction identified an epigenetic crosstalk. The dashed purple edge between LCOR and EHMT2 was inferred by the graph from surrounding structural evidence, not directly observed in any single dataset.

The Drug Development Landscape in the PRC2 cluster

To assess actionability, the team mapped each gene in the PRC2 cluster to its current stage in drug development:

This druggability assessment drew on the Base KG's integration of ChEMBL clinical stage data and UniProt protein domain annotations (Layer 1), cross-referenced with the co-essentiality and link prediction outputs from Layer 3.

Gene	Inhibitor(s)	Clinical Stage
EZH2	Tazemetostat	Approved
EED	MAK683 (Novartis)	Phase I/II
EHMT2	UNC0642, A366	Pre-clinical
REST	X5050	Research only
JARID2	Via PRC2 inhibitors	No direct inhibitor
SUZ12	Via EZH2 inhibitors	No direct inhibitor

KEY FINDING

EHMT2 is the only pre-clinical marker in the PRC2 cluster with demonstrated active drug chemistry. Both inhibitors (UNC0642 and A366) are in use in cell-culture experiments. REST also appeared but its inhibitor has shown no development progress in over a decade. EHMT2 represents the primary open opportunity in a validated chromatin axis.

Biological and Clinical Validation

Validating Neuroendocrine Specificity-

To verify this predicted link was not a computational artifact of generic prostate biology, the Knowledge Graph conducted a cross-cohort gene expression correlation using TCGA data. The results confirmed strict neuroendocrine specificity: the LCOR-EHMT2 link showed a significant positive correlation in neuroendocrine tumors PCPG (r = 0.287) and a stark negative correlation in prostate adenocarcinoma (PRAD r = -0.292), statistically validated by a Fisher z-test ($p = 4.40 \times 10^{-9}$).

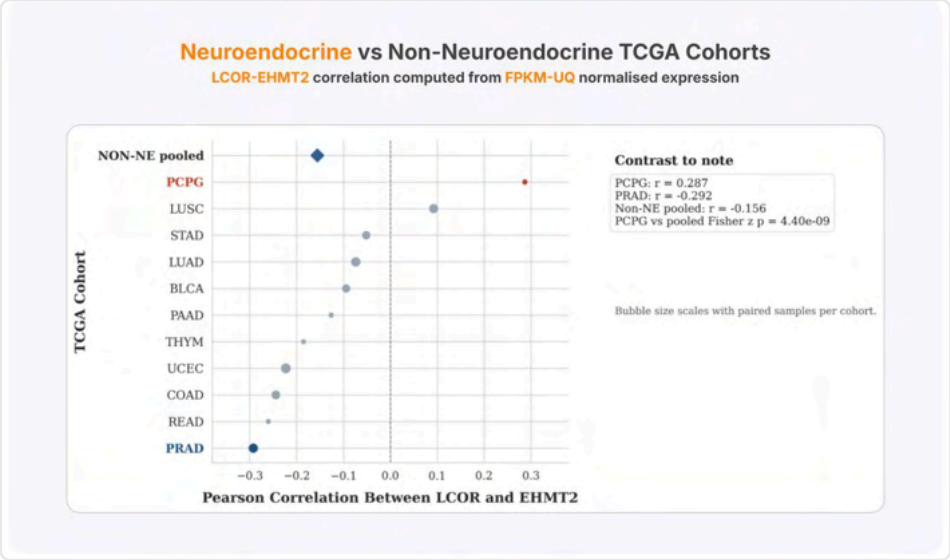


Figure 5. LCOR–EHMT2 expression correlation across TCGA cohorts. Positive correlation in the neuroendocrine cohort (PCPG), negative in non-neuroendocrine cancers including prostate adenocarcinoma (PRAD).

EHMT2:
External Literature Convergence

This interaction, predicted entirely by the Knowledge Graph, is firmly corroborated by external literature. Independent studies in adjacent cancers demonstrate that targeting EHMT2 (G9a) reverses EGFR-TKI resistance in non-small cell lung cancer through PTEN/AKT regulation, and its suppression enhances NK cell-mediated anti-tumor immunity via TGF-β1 suppression.

The graph did not merely predict a novel link; it predicted an actionable link consistent with published functional evidence from independent biological contexts.

Future Directions

The framework demonstrated in this case study generalizes to any disease where conventional targets have failed and biological signals are highly fragmented. Active development is focused on:

Expansion across oncology indications: Applying the framework to additional cancers characterized by data scarcity, treatment resistance, or lineage plasticity.

Drug repurposing: Identifying therapeutic opportunities by connecting shared mechanisms across diseases and treatment contexts.

Pathway crosstalk discovery: Surfacing additional regulatory dependencies that are invisible in single-dataset analyses.

Resistance mechanism modeling: Characterizing adaptive signaling and resistance pathways early in disease progression, before clinical manifestation.

This work was presented at the American Association for Cancer Research (AACR) Annual Meeting in April 2026 as Poster: “Knowledge graph driven insights and drug repurposing opportunities for neuroendocrine prostate cancer.”

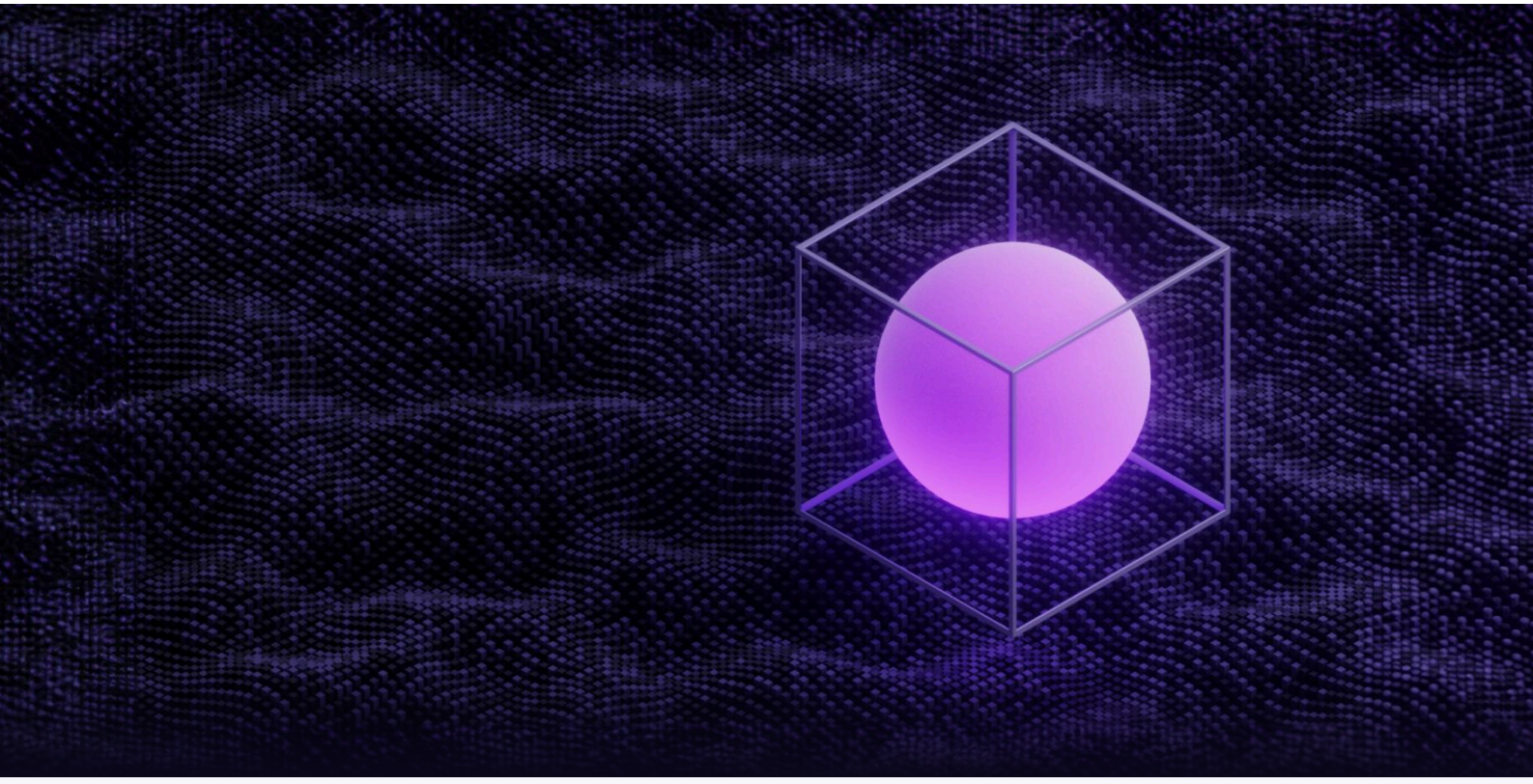
Conclusion

The regulatory crosstalks that drive treatment resistance, lineage plasticity, and immune evasion in cancer are rarely published as discrete gene-disease pairs. They are scattered across independent screens, expression datasets, and clinical cohorts - sources that flat databases cannot integrate and that co-occurrence models cannot infer.

By constructing a causal knowledge graph around the specific biology of NEPC, this analysis successfully bypassed literature limitations. Moving from an undruggable master driver (MYCN) to an actionable pre-clinical target (EHMT2) demonstrates that even in severely data-constrained environments, mechanistically grounded computational discovery can uncover viable, novel clinical opportunities that standard tools like Open Targets completely miss.

The finding was independently supported by published functional evidence from adjacent cancers and confirmed as neuroendocrine-specific through cross-cohort correlation analysis. The knowledge graph did not merely surface a candidate gene. It predicted a regulatory interaction that explained how a validated chromatin axis could be targeted through a mechanism that no single dataset contained.

This framework generalizes beyond NEPC to any indication where critical regulatory crosstalks remain hidden across fragmented datasets, and where discovering those crosstalks is the path to actionable biology.



References

1. Nelson MR, Tipney H, Painter JL, et al. The support of human genetic evidence for approved drug indications. Nature Genetics. 2015;47(8):856–860.
2. King EA, Davis JW, Degner JF. Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval. PLoS Genetics. 2019;15(12):e1008489.
3. Paul SM, Mytelka DS, Dunwiddie CT, et al. How to improve R&D productivity: the pharmaceutical industry's grand challenge. Nature Reviews Drug Discovery. 2010;9(3):203–214.
4. Wang L, Dong X, Ren Y, et al. Targeting EHMT2 reverses EGFR-TKI resistance in NSCLC by epigenetically regulating the PTEN/AKT signaling pathway. Cell Death & Disease. 2018;9:129.
5. AACR 2026 Poster #6873: Knowledge graph driven insights and drug repurposing opportunities for neuroendocrine prostate cancer. Elucidata.

Glossary

NEPC
Neuroendocrine Prostate Cancer. An aggressive, treatment-resistant subtype arising through lineage plasticity.

Lineage plasticity
The ability of cancer cells to change cell identity to evade therapy.

MYCN
A transcription factor and master regulator of NEPC. Currently undruggable.

PRC2 complex
Polycomb Repressive Complex 2. A chromatin-modifying complex central to lineage plasticity.

EHMT2 (G9a)
Euchromatic Histone Lysine Methyltransferase 2. The druggable chromatin regulator identified through this analysis.

Co-essentiality
A quantitative measure of functional dependency between genes operating within the same complex or pathway.

Link prediction
An ML technique that infers novel relationships between graph entities based on structural patterns.

Polly Scout
Elucidata's AI-powered dataset discovery tool for querying 1M+ public biomedical records.

Base KG
The foundational layer of Polly KG, integrating 20+ curated public data sources.

AR
Androgen Receptor. The primary therapeutic target in standard prostate cancer, lost during the NEPC transition.

ARPI
Androgen Receptor Pathway Inhibitor (e.g., enzalutamide). The therapy class whose selective pressure drives NEPC emergence.

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