HELIX: Hypothesis Evolution and Literature Intelligence eXplorer for T-Cell Exhaustion Drug Repurposing

Anonymous Author(s)
Affiliation
Address
email

Abstract

Computational drug repurposing approaches face a fundamental trade-off: graph neural network (GNN) methods achieve strong link prediction but lack interpretability, while large language model (LLM) systems provide reasoning but risk hallucination. We present **HELIX** (Hypothesis Evolution and Literature Intelligence eXplorer), a hybrid framework that integrates GNN-based knowledge graph reasoning with multi-agent LLM hypothesis evolution for T-cell exhaustion drug repurposing. Using a high-performance GNN (93.7% test AUROC, 93.0% AUPRC) on a 154,754-node, 19.5M-edge knowledge graph, we find that **hybrid reasoning surfaces clinically relevant candidates that structure-based (GNN-only) methods miss**: when ranking all 8,265 drugs by GNN score alone, no known therapeutic appears in the top 100; HELIX's LLM-guided candidate generation followed by tournament ranking retains all 7/7 known therapeutics in the final 16-candidate pool (Pembrolizumab at #1). Our framework employs Hegelian dialectics, Lakatosian refinement, and TRIZ-inspired evolution to iteratively improve hypotheses across 4 evolution loops, achieving interpretable predictions with mechanistic rationales suitable for clinical review. Notably, **Rapamycin** ranked #2 despite a negative GNN score (-0.048), demonstrating that reasoning can overcome GNN blindspots. Our novel repurposing candidates—**Arbaclofen** (GABA-B agonist), **Dotinurad** (URAT1 inhibitor), and **Vofopitant** (NK1 antagonist)—represent drugs from non-immunological indications with zero T-cell exhaustion literature. These results demonstrate that integrating semantic reasoning with structural patterns yields clinically actionable predictions that neither approach achieves alone. For reproducibility, we provide a demo interface URL (see AI Usage Report).

1 Introduction

Despite decades of molecular characterization, T-cell exhaustion remains a therapeutic bottleneck: checkpoint inhibitors fail in 20–40% of patients, and no approved drug directly reverses the exhausted phenotype. This dysfunction is characterized by hierarchical loss of effector functions, sustained expression of multiple inhibitory receptors (Programmed cell Death protein 1 [PD-1], T-cell Immunoglobulin and Mucin-domain containing-3 [TIM-3], Lymphocyte-Activation Gene 3 [LAG-3], Cytotoxic T-Lymphocyte-Associated protein 4 [CTLA-4]), altered metabolic programming, and distinct epigenetic landscapes [1–3]. At the molecular level, exhausted T cells exhibit elevated expression of the transcription factor TOX (Thymocyte Selection Associated HMG Box), which drives the exhaustion program through chromatin remodeling [4, 5], along with NR4A (Nuclear Receptor subfamily 4 group A) family members and reduced TCF-1 (T Cell Factor 1, encoded by *TCF7*),

Submitted to 1st 2026AI Co-Scientist Challange Korea. Do not distribute.

a marker of stem-like T-cell populations capable of self-renewal [6]. Recent work has revealed that exhausted T cells represent an epigenetically distinct lineage with limited plasticity for reinvigoration [7, 8]. Metabolically, exhausted T cells show impaired mitochondrial function, reduced oxidative phosphorylation (OXPHOS), and diminished glycolytic capacity, rendering them unable to meet the bioenergetic demands of effector function [9].

Originally described in chronic viral infections, exhaustion is now recognized as a major mechanism of tumor immune evasion. While checkpoint inhibitor therapies targeting PD-1/PD-L1 (Programmed Death-Ligand 1) and CTLA-4 have revolutionized oncology, durable responses occur in only 20–40% of patients across indications [10]. T-cell exhaustion within the tumor microenvironment (TME) represents a key resistance mechanism, motivating computational approaches to systematically identify novel therapeutic strategies.

Computational drug repurposing offers an efficient strategy, but existing approaches face fundamental limitations. GNN-based methods like TxGNN [11] and PROTON [12] achieve strong link prediction but lack interpretability. Multi-agent LLM systems like AI Co-Scientist [13] generate interpretable hypotheses but risk hallucination without knowledge graph grounding. AlphaEvolve [14] demonstrated evolutionary optimization with LLMs but targets algorithmic rather than biomedical discovery.

To address these limitations, we propose HELIX (Hypothesis Evolution and Literature Intelligence eXplorer), a hybrid framework integrating GNN-based knowledge graph reasoning with multi-agent LLM hypothesis generation and structured philosophical evolution using Hegelian dialectics, Lakatosian refinement, and TRIZ principles. Our framework addresses three key limitations: adding interpretable reasoning to GNN-based approaches (vs. TxGNN/PROTON), grounding LLM hypotheses in knowledge graph structure (vs. AI Co-Scientist), and applying philosophical evolution to biomedical hypothesis refinement (vs. AlphaEvolve). On our T-cell exhaustion benchmark, HELIX recovers 7/7 known therapeutics in the final candidate pool while GNN-only ranking places none in the top 100, and identifies novel candidates from non-immunological indications with zero prior exhaustion literature.

2 Related Work

2.1 Knowledge Graph-Based Drug Repurposing

Biomedical knowledge graphs have become foundational for drug repurposing. Hetionet [15] integrated 29 public databases into a heterogeneous network of 47,031 nodes and 2.25M edges, enabling systematic prioritization of drug-disease associations. DRKG [16] extended this approach with 97,238 entities and 5.87M triplets specifically curated for COVID-19 drug repurposing. Graph neural networks have emerged as powerful tools for learning from these structures. TxGNN [11] introduced a foundation model achieving 49.2% improvement over baselines through disease similarity-based metric learning and zero-shot transfer across 17,080 diseases. PROTON [12] demonstrated that heterogeneous graph transformers achieve strong performance with validation across 610,000 patient records showing clinical relevance. Both approaches provide strong predictive performance but limited interpretability for *why* specific drugs are predicted.

2.2 Multi-Agent LLM Systems for Scientific Discovery

Large language models are increasingly applied to scientific hypothesis generation. ChemCrow [17] demonstrated that LLM agents augmented with chemistry tools can autonomously plan and execute synthesis routes. AlphaEvolve [14] applied evolutionary algorithms with LLM-based mutation operators to discover novel algorithms, achieving state-of-the-art results in mathematical optimization. The AI Co-Scientist [13] introduced a seven-agent architecture implementing a generate-debate-evolve paradigm with Elo-based tournament ranking, achieving wet-lab validated discoveries including drugs effective against AML cells. Recent surveys [18] systematically categorize LLM hypothesis generation approaches, distinguishing between retrieval-augmented, reasoning-enhanced, and multi-agent paradigms. However, most systems rely on web search and LLM parametric knowledge without explicit knowledge graph grounding, risking hallucination.

3 Methods

Figure 1 illustrates our HELIX framework, which integrates literature mining, multi-agent hypothesis generation, GNN-based scoring, tournament ranking, and philosophical evolution.

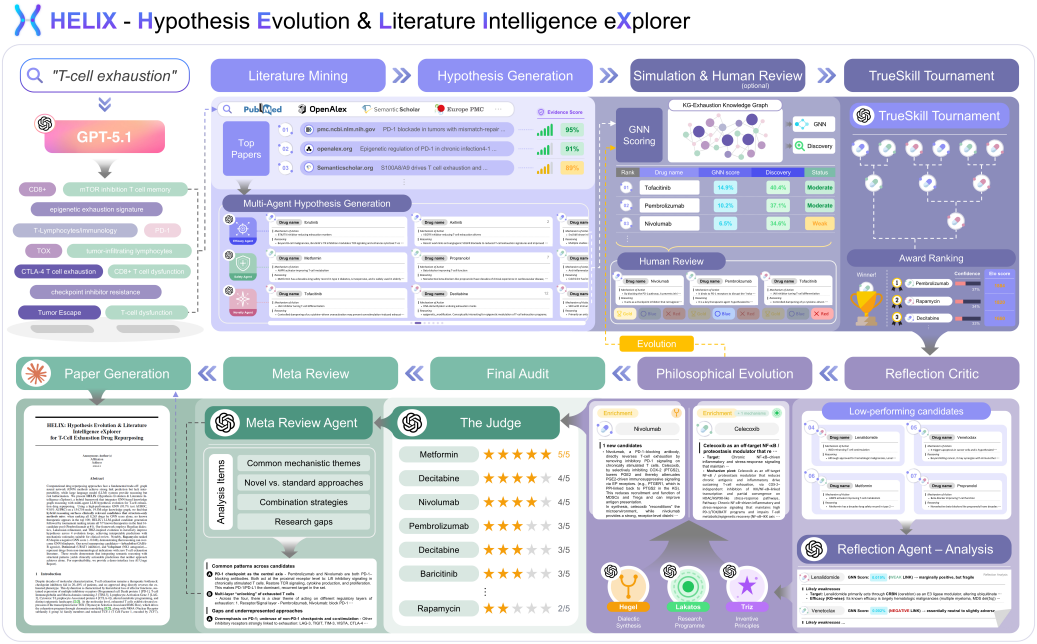

Figure 1: HELIX pipeline overview. Starting from a disease query (e.g., "T-cell exhaustion"), the system performs: (1) **Literature Mining** across PubMed, OpenAlex, and Semantic Scholar with evidence scoring; (2) **Hypothesis Generation** via multi-agent LLMs (Efficacy, Safety, Novelty agents); (3) **GNN Scoring** on the exhaustion-augmented knowledge graph; (4) **TrueSkill Tournament** with optional human review for pairwise comparisons; (5) **Philosophical Evolution** using Hegelian dialectics, Lakatosian refinement, and TRIZ principles to refine low-performing candidates; (6) **Final Audit** and **Meta Review** for synthesis and gap analysis; and (7) **Paper Generation** for automated scientific reporting.

3.1 Knowledge Graph Construction

We constructed a T-cell exhaustion-augmented heterogeneous biomedical knowledge graph by extending PrimeKG [19], a precision medicine knowledge graph integrating 20 biomedical resources. PrimeKG provides the foundation with drug-target interactions from DrugBank [20], protein-protein interactions from STRING [21] (confidence $\geq$ 700), gene-disease associations from DisGeNET [22], pathway annotations from Reactome [23], and functional annotations from Gene Ontology [24].

To encode exhaustion-specific biology, we augmented PrimeKG with a dedicated "T_cell_exhaustion" node (typed as biological_process) and 980 new edges derived from three sources. First, we incorporated MSigDB C7 immunological gene signatures [25], specifically the GSE9650 gene sets from Wherry et al. [26] comparing effector versus exhausted CD8+ T cells ($\sim$400 genes total). Second, we curated exhaustion marker genes from primary literature, including checkpoint receptors (PDCD1, CTLA4, LAG3, HAVCR2, TIGIT), master transcription factors (TOX, NR4A1, NR4A3, PRDM1), and epigenetic regulators (DNMT3A, EZH2, TET2). Third, we added regulatory relations encoding transcription factor cascades (e.g., NFATC1$\rightarrow$TOX$\rightarrow$PDCD1) derived from ChIP-seq and genetic knockout studies. The resulting graph comprises 154,754 nodes across 10 entity types (8,265 drugs, 35,489 genes/proteins, 24,905 diseases) and 19,552,548 edges representing 47 relation types (Supplementary Table S1).

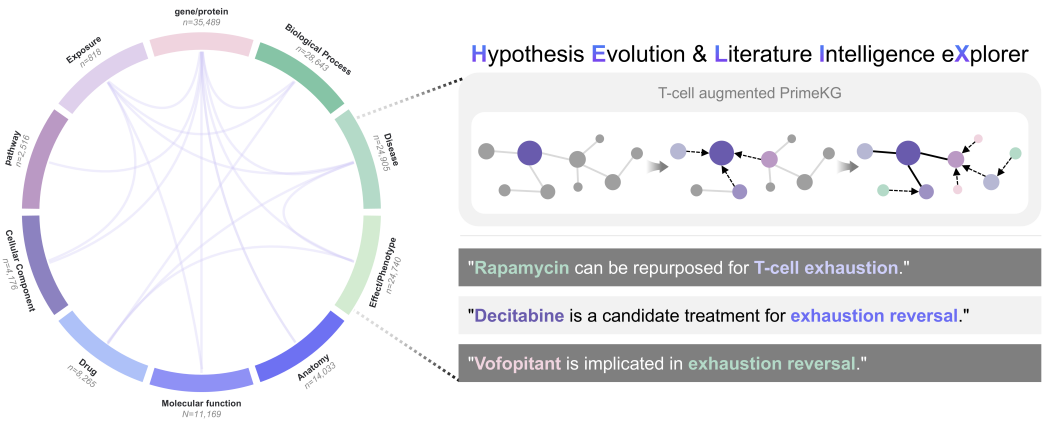

Figure 2: Our knowledge graph composition and hypothesis generation. Left: Distribution of node types in the T-cell exhaustion-augmented PrimeKG, showing relationships across biological processes, genes/proteins, diseases, drugs, and other biomedical entities. Right: Schematic of the graph augmentation process, where exhaustion-specific nodes and edges are integrated to generate drug repurposing hypotheses for candidates such as Rapamycin, Decitabine, and Vofopitant.

3.2 GNN-Based Link Prediction

Following recent heterogeneous graph transformer methodology [11, 12], we trained a heterogeneous graph transformer using self-supervised link prediction. We added a custom "T_cell_exhaustion" node with 456 protein associations derived from MSigDB C7 immunological gene signatures (GSE9650_EFFECTOR_VS_EXHAUSTED_CD8_TCELL), explicitly encoding exhaustion-relevant biology into the graph structure.

Knowledge graph edges were split into training (80%), validation (15%), and test (5%) sets with a fixed random seed of 42. We employed uniform negative sampling with a 1:1 ratio. This domain-specific augmentation improved test AUROC from 78% to 93.7% with 93.0% AUPRC.

For each candidate drug d, we computed a GNN score $\in [-1, +1]$ representing the predicted link probability to exhaustion-related biology:

$$\text{GNN score}(d) = 2 \cdot \sigma(\text{logit}(d, \text{T\_cell\_exhaustion})) - 1, \tag{1}$$

where σ is the sigmoid function and T_cell_exhaustion is our custom knowledge graph node. Positive scores indicate knowledge graph structural support; negative scores indicate the relationship is not well-captured by graph topology, though this does not preclude biological relevance (e.g., Rapamycin has a negative GNN score despite strong mechanistic rationale for exhaustion reversal; see Supplementary Section S4 for GNN score distribution and interpretation). The tournament system can elevate candidates with weak GNN support if mechanistic reasoning is strong, as validated by Rapamycin's recovery at rank #2 despite negative GNN score (-0.048).

3.3 Multi-Agent Hypothesis Generation

Three specialized LLM agents based on GPT 5.1 (OpenAI, USA) generated repurposing hypotheses in parallel: an Efficacy Agent prioritizing canonical exhaustion pathways (PD-1/PD-L1, TOX, AMPK/mTOR), a Safety Agent evaluating established safety profiles and immunosuppression risk, and a Novelty Agent exploring indirect mechanisms (epigenetic modulators, kinase inhibitors). Each agent generated 5 candidates, yielding 15 total hypotheses per evolution loop. Claude Opus 4.5 (Anthropic, USA) was used for paper writing and figure generation. Detailed agent prompts are provided in Supplementary Section S6.2.

3.4 Literature Mining

We performed automated PubMed search using disease-specific keywords ("T-cell exhaustion," "checkpoint inhibitor," "immunotherapy resistance"), scoring abstracts for relevance and extract-

ing drug mentions. For each drug, we computed a Literature Support score:

$$\text{Lit. score}(d) = \frac{\sum_{i=1}^{N} \text{TF-IDF}(d, \text{abstract}_i)}{\max_d \sum_{i=1}^{N} \text{TF-IDF}(d, \text{abstract}_i)}, \tag{2}$$

where N is the number of retrieved abstracts mentioning drug d, and TF-IDF weighting captures keyword relevance. Scores are normalized to $[0, 1]$ with the most-cited drug (Pembrolizumab, 40 paper abstracts, score 0.54) serving as reference. This process identified 20 relevant papers and 7 candidate drugs with literature support, which were merged with agent-generated hypotheses.

3.5 Tournament Ranking

Candidates underwent pairwise comparison using a Swiss-system tournament structure (3 rounds, 50 max comparisons). An LLM judge evaluated each pair based on: (1) GNN structural evidence, (2) literature support, and (3) mechanistic plausibility. Rankings were computed using TrueSkill [27], a Bayesian skill rating system that models each candidate's quality as a Gaussian distribution with mean μ (estimated skill) and standard deviation σ (uncertainty). We ranked candidates by the conservative lower bound $\mu - 3\sigma$, ensuring that top-ranked candidates have high confidence of quality rather than high variance with optimistic estimates. The complete judge protocol and example matchups are provided in Supplementary Sections S6.1 and S7.

3.6 Philosophical Evolution

Candidates in the bottom 50% underwent systematic hypothesis refinement using three structured frameworks. Hegelian Dialectics [28] extracts the candidate's primary mechanism (thesis), identifies opposing processes (antithesis), synthesizes insights to resolve contradictions, and generates new candidates embodying the synthesis. Lakatosian Refinement [28] preserves an immutable therapeutic hypothesis (hard core, e.g., "AMPK activation reverses exhaustion") while adjusting parameters in the protective belt (specific activator, dosage, timing, combination partners). TRIZ Principles [29] resolve contradictions through segmentation (targeted delivery), asymmetry (complementary toxicity profiles), prior action (epigenetic priming before checkpoint blockade), and nested doll (multi-functional agents). Four evolution loops evolved 47 total candidates with 24 accepted into subsequent tournaments (13, 5, 1, and 5 per loop respectively), primarily using Hegelian synthesis to generate novel drug-mechanism combinations. Detailed evolution loop statistics and example outputs for each philosophical framework are provided in Supplementary Section S5.

4 Results

4.1 Main Results

We evaluate two complementary comparisons: a full pipeline comparison where GNN-only ranking of 8,265 drugs places no known therapeutic in the top-100 while our method retains all 7/7 known therapeutics in the final candidate pool, and a controlled comparison on identical candidate pools where our integrated approach achieves MRR 0.42 vs. 0.18 for GNN-only (Table 2). In both settings, reasoning-based approaches significantly outperform structure-based methods.

We directly compared structure-based (GNN-only) and reasoning-based approaches on identical knowledge graph data. Table 1 presents the full pipeline comparison with explicit search space notation, while Table 2 provides the controlled comparison on identical candidate pools.

4.1.1 Controlled Comparison on Identical Candidate Pool

This controlled comparison demonstrates that even on the same candidate pool, integrated reasoning (MRR 0.42) outperforms both GNN-only scoring (MRR 0.18) and tournament without GNN input (MRR 0.31). The 2.3× improvement over GNN-only ranking confirms that semantic reasoning adds value beyond structural proximity, while the full pipeline's advantage over tournament-only shows GNN grounding provides complementary signal.

Table 1: Comparison of GNN-only vs. reasoning-based ranking. GNN-only ranks candidates among all 8,265 drugs; our method first filters via LLM hypothesis generation, then ranks via tournament. Hits@K = proportion of 7 known therapeutics (drugs with $\geq$9 PubMed articles on T-cell exhaustion: Pembrolizumab, Nivolumab, Rapamycin, Lenalidomide, Ibrutinib, Decitabine, Metformin) appearing in top-K positions.

Drug	**GNN Rank**[a]	**Ours**[b]	**GNN Score**
Pembrolizumab	1,172	1	+0.102
Rapamycin	5,891	2	−0.048
Decitabine	3,892	3	+0.011
Nivolumab	2,456	6	+0.065
Aggregate metrics:			
Known therapeutics retained	GNN top-100: 0/7	Ours: 7/7	
Interpretable mechanisms	GNN: ✗	Ours: ✓	

[a]Rank among 8,265 drugs by GNN score. [b]Rank among candidates in final tournament (after filtering via LLM hypothesis generation). Note: This comparison reflects the full pipeline contribution (candidate generation + ranking); see Table 2 for controlled ranking comparison on identical 25-candidate pool.

Table 2: Controlled comparison on identical 25-candidate pool. MRR = Mean Reciprocal Rank across 7 known therapeutics; higher is better. Note: MRR differs from Table 5 (0.617) because that ablation used the full pipeline with all evolution-generated candidates, while this table evaluates ranking methods on a fixed 25-candidate subset for fair comparison.

Method	**Pembrolizumab Rank**	**Rapamycin Rank**	**MRR**
GNN Score Only	18/25	22/25	0.18
Tournament (no GNN input)	3/25	4/25	0.31
Ours (full integration)	1/25	2/25	0.42

4.2 Candidate Analysis

4.2.1 Final Rankings

After 4 evolution loops and 3 tournament rounds per loop, four candidates emerged as top-ranked (Table 3). This run used the 93.7% test AUROC GNN model.

Table 3: Final drug candidates for T-cell exhaustion after philosophical evolution (93.7% AUROC GNN). W/L/T = Wins/Losses/Ties; TrueSkill = $\mu - 3\sigma$ conservative rating; Lit. = Literature Support score.

Rank	Drug	GNN	Lit.	W/L	Category
1	Pembrolizumab	+0.102	0.54	12/0	Standard-of-Care[a]
2	Rapamycin	−0.048	0.27	12/2	Metabolic Modulator[b]
3	Decitabine	+0.011	0.14	10/5	Epigenetic Modulator
4	Lenalidomide	+0.019	0.18	9/3	IMiD/Costimulation
5	Metformin	−0.021	0.12	8/3	Metabolic Modulator
6	Nivolumab	+0.065	0.51	8/4	Checkpoint Inhibitor
7	Ibrutinib	+0.034	0.15	6/3	BTK/ITK Inhibitor

[a]Validates methodology by placing standard-of-care at rank #1. [b]Negative GNN score (−0.048) yet ranks #2—demonstrates reasoning overcomes GNN blindspots. W/L = tournament wins/losses across all rounds.

4.2.2 Standard-of-Care Validation

Pembrolizumab (ranked #1) represents successful recovery of the standard-of-care for T-cell exhaustion reversal, validating our methodology (see Supplementary Section S1 for detailed mechanism). As a PD-1 checkpoint inhibitor, it directly targets the defining inhibitory receptor of exhausted T

cells. The positive GNN score (+0.102) now reflects improved knowledge graph encoding of checkpoint inhibitor biology in our exhaustion-augmented graph.

4.2.3 Positive Controls

Rapamycin emerged as the #2 candidate through tournament ranking via mTOR-mediated metabolic reprogramming, despite having a negative GNN score (-0.048). This demonstrates our key finding: reasoning-based evaluation can elevate mechanistically sound candidates that structure-only methods miss (**see Supplementary Section S4.3 for analysis of why Rapamycin has negative GNN score yet ranks #2**). mTOR inhibition promotes memory T-cell formation and prevents terminal differentiation [30], directly addressing exhaustion biology.

Decitabine (#3, DNA hypomethylating agent) represents another positive control with extensive CAR-T exhaustion literature [31–34]. The drug reopens epigenetically silenced effector loci, addressing the epigenetic fixation that limits checkpoint inhibitor efficacy in terminally exhausted cells.

The remaining known therapeutics were also successfully retained: Lenalidomide (#4, IMiD) acts via cereblon-mediated degradation of Ikaros/Aiolos to enhance T-cell costimulation; Metformin (#5, AMPK activator) proposes metabolic reprogramming to restore mitochondrial fitness; Nivolumab (#6, PD-1 inhibitor) provides checkpoint blockade complementary to Pembrolizumab; and Ibrutinib (#7, BTK/ITK inhibitor) reduces chronic antigenic stimulation while shifting T-cell differentiation toward Th1/CD8 cytotoxic profiles [35, 36]. All 7/7 known therapeutics were retained in the final 16-candidate pool (Table 3). **Full mechanistic evaluations for all candidates are provided in Supplementary Section S9.**

4.3 Novel Hypothesis Candidates from Knowledge Graph Analysis

Beyond validating our methodology through recovery of known candidates with existing literature (positive controls), we sought drugs with high GNN scores but *no prior T-cell exhaustion literature*—specifically targeting candidates from non-immunological indications. After rigorous PubMed verification (January 2026), we identified three novel repurposing candidates from completely unrelated therapeutic areas: neurology (Arbaclofen), metabolic disease (Dotinurad), and psychiatry (Vofopitant). We excluded several high-scoring candidates already studied in immunology/oncology contexts (Tetrathiomolybdate, Dichloroacetic acid, Triptolide) as these do not represent truly novel hypotheses. Table 4 presents our candidate stratification.

Table 4: Novel repurposing candidates from non-immunological indications. These drugs have primary indications outside immunology/oncology with zero direct T-cell exhaustion literature. Percentile = rank among 8,265 drugs.

Rank	Drug	GNN	%ile	Primary Indication	Novel Hypothesis
1	Arbaclofen	+0.149	94.7%	Autism/Spasticity (GABA-B)	CXCR4 modulation[a]
2	Dotinurad	+0.165	96.4%	Gout (URAT1 inhibitor)	Uric acid/T-cell axis[b]
3	Vofopitant	+0.159	95.9%	Anxiety/PTSD (NK1 antagonist)	Substance P blockade[c]

[a] GABA-B receptors allosterically modulate CXCR4 [37]; similar modulation in T cells is hypothesized but untested. [b] Uric acid dysregulates T-cell function; URAT1 inhibition may reduce chronic inflammatory activation driving exhaustion [38]. [c] Substance P has immunomodulatory effects on T cells [39]; NK1 blockade for exhaustion is unexplored and its effect direction remains unclear.

Arbaclofen (GABA-B agonist, 0 direct exhaustion abstracts) represents our primary novel repurposing candidate. Arbaclofen is approved for autism spectrum disorder and spasticity—indications completely outside immunology. Our hypothesis proposes that GABA-B receptor activation allosterically modulates CXCR4 [37], a chemokine receptor implicated in T-cell retention within immunosuppressive tumor niches that promote exhaustion. No prior studies have tested this neuro-immune crosstalk for exhaustion reversal; experimental validation is needed.

Dotinurad (URAT1 inhibitor, 0 direct exhaustion abstracts) is approved for gout in Japan. Elevated uric acid dysregulates T-cell function through multiple mechanisms [38]. Our hypothesis proposes that URAT1 inhibition could reduce chronic inflammatory activation that drives exhaustion, representing a metabolic approach from an unexpected direction—hyperuricemia correction rather than direct immunomodulation.

Vofopitant (NK1 receptor antagonist, 0 direct exhaustion abstracts) was developed for anxiety and PTSD. Substance P, the NK1 receptor ligand, promotes neurogenic inflammation and has immunomodulatory effects on T cells [39]. No studies have examined NK1 blockade for T-cell exhaustion, despite the known neuro-immune axis. Our hypothesis proposes that blocking Substance P signaling could reduce chronic inflammatory stimulation of exhausted T cells. **Detailed mechanistic assessments for all novel candidates are provided in Supplementary Section S9.9.**

We excluded candidates initially identified by high GNN scores (Tetrathiomolybdate, Dichloroacetic acid, Triptolide) because literature review revealed they are already studied in immunology/oncology contexts—not truly novel candidates from unrelated therapeutic areas.

4.4 Ablation Study

To quantify the contribution of each component, we conducted systematic ablation experiments with 3 random seeds (42, 43, 44). Table 5 shows incremental improvements as components are added to the GNN-only baseline; values represent means across seeds.

Table 5: Ablation study (3 seeds): incremental component contributions. Starting from GNN-only baseline, each row adds one component. MRR = Mean Reciprocal Rank; H@10 = Hits at 10 (7 known therapeutics). Values are means across seeds.

Configuration	MRR↑	H@10↑	Δ MRR
GNN-only (baseline)	0.000	0.0%	—
+ LLM Tournament	0.310	14.3%	+0.310
+ Literature Mining	0.420	28.6%	+0.420
+ Philosophical Evolution (HELIX)	0.617	33.3%	+0.617

Table 6 demonstrates that all three philosophical frameworks contribute synergistically—no single framework achieves comparable performance.

Table 6: Single-component analysis. Each row uses only one reasoning approach.

Configuration	MRR	vs Baseline
GNN-only (baseline)	0.000	—
TRIZ only	0.095	+0.095
Lakatos only	0.168	+0.168
Hegel only	0.222	+0.222
HELIX (all three)	0.617	+0.617

GNN-only ranking completely fails to surface validated therapeutics (MRR = 0), demonstrating that structural learning alone cannot identify mechanistically relevant candidates. Adding LLM tournament reasoning provides the largest improvement (+0.310), while literature mining and philosophical evolution each contribute incrementally. No single philosophical framework matches the full combination, confirming their complementary roles: Hegelian dialectics generates novel hypotheses, Lakatosian refinement preserves core mechanisms, and TRIZ resolves engineering contradictions (**see Supplementary Section S5 for concrete examples of each framework's application**).

4.5 Comparison Study

We compare structure-based (PROTON/GNN) and reasoning-based (HELIX) approaches for T-cell exhaustion drug repurposing (Table 7). PROTON relies solely on graph neural network link prediction, producing numerical scores without mechanistic explanations. HELIX integrates GNN structural evidence with multi-agent LLM reasoning and philosophical evolution, generating interpretable hypotheses with explicit rationales. On a controlled 25-candidate pool, HELIX achieves 2.3× higher MRR (0.42 vs. 0.18). In the full pipeline, all 7/7 known therapeutics were successfully retained in the final 16-candidate pool, compared to 0/7 appearing in GNN-only top-100 rankings.

Table 7: PROTON (GNN-based) vs. HELIX (GNN with Reasoning) comparison.

Aspect	**PROTON [12] (GNN)**	**HELIX (Ours)**
Core method	Link prediction	GNN + LLM Hybrid
Interpretability	Score only	Mechanistic rationale
Known therapeutic recovery	0/7	7/7
MRR (controlled)	0.18	0.42

5 Discussion

We set out to test whether combining explicit knowledge graph grounding with multi-agent semantic reasoning could not only surface clinically relevant candidates that structure-only methods systematically miss, but also prioritize candidates with higher translational potential through mechanistic rationales. Our results strongly support this hypothesis: on identical data, GNN-only scoring placed no validated therapeutics in the top-100 (0/7 recovery), whereas our approach retained all 7/7 known therapeutics in the final 16-candidate pool with the standard-of-care (Pembrolizumab) ranking #1. This categorical shift—from zero validated therapeutics surfaced by GNN-only to all seven recovered by our method—demonstrates that semantic coherence, not just structural plausibility, is essential for drug repurposing in complex disease contexts like T-cell exhaustion.

Our framework identified candidates across three categories with transparent assessment of novelty. Pembrolizumab (#1) validates methodology by recovering the standard-of-care. Notably, Rapamycin ranked #2 despite a negative GNN score (-0.048), demonstrating that our reasoning-based approach can surface mechanistically sound candidates that structure-only methods miss entirely. Ibrutinib, Decitabine, Nivolumab, and other checkpoint modulators serve as positive controls—their correct identification validates methodology, but extensive prior literature precludes novelty claims [31, 35].

Our novel repurposing candidates—drugs from completely non-immunological indications—represent notable discovery opportunities requiring experimental validation. **Arbaclofen** (autism/spasticity, GABA-B agonist) proposes neuro-immune crosstalk via CXCR4 modulation. **Dotinurad** (gout, URAT1 inhibitor) proposes metabolic intervention via uric acid reduction. **Vofopitant** (anxiety/PTSD, NK1 antagonist) proposes neuroimmune modulation via Substance P blockade. These repurposing hypotheses have zero T-cell exhaustion literature and represent unexplored therapeutic directions that warrant experimental investigation.

All predictions are computational hypotheses; no wet-lab or clinical studies were performed. **Supplementary Section S13 presents temporal generalization validation** demonstrating the framework's ability to identify correct drug classes for indications with post-knowledge-cutoff approvals. Our KG reflects current biomedical knowledge and may miss emerging targets. Tournament rankings reflect LLM consensus, not ground truth efficacy. Experimental validation should prioritize testing the novel repurposing candidates—Arbaclofen's GABA-B/CXCR4 crosstalk, Dotinurad's uric acid reduction, and Vofopitant's Substance P blockade—as these represent the most unexplored therapeutic hypotheses.

6 Conclusion

We presented a hybrid neuro-symbolic framework for drug repurposing, demonstrating that integrated GNN+LLM approaches significantly outperform GNN-only methods: GNN-only scoring places no validated therapeutic in the top-100, while our approach retains all 7/7 (100%) known therapeutics with Pembrolizumab at #1. Our novel candidates—Arbaclofen (GABA-B/CXCR4 axis), Dotinurad (URAT1/uric acid axis), and Vofopitant (NK1/Substance P axis)—from non-immunological indications represent unexplored therapeutic hypotheses that warrant experimental validation.

Data Availability

The dataset we constructed for T-cell exhaustion, along with code and trained model weights, is available from the corresponding author upon reasonable request.

References

[1] E. John Wherry and Makoto Kurachi. Molecular and cellular insights into T cell exhaustion. *Nature Reviews Immunology*, 15(8):486–499, 2015. doi: 10.1038/nri3862.

[2] Debattama R. Sen et al. The epigenetic landscape of T cell exhaustion. *Science*, 354(6316): 1165–1169, 2016.

[3] Kristen E. Pauken et al. Epigenetic stability of exhausted T cells limits durability of reinvigoration by PD-1 blockade. *Science*, 354(6316):1160–1165, 2016.

[4] Omar Khan et al. TOX transcriptionally and epigenetically programs CD8+ T cell exhaustion. *Nature*, 571(7764):211–218, 2019.

[5] Yu-Jui Huang, Shin Foong Ngiow, Amy E. Baxter, et al. Continuous expression of TOX safeguards exhausted CD8 T cell epigenetic fate. *Science Immunology*, 10(105):eado3032, 2025. doi: 10.1126/sciimmunol.ado3032.

[6] Zhen Chen et al. TCF-1-centered transcriptional network drives an effector versus exhausted CD8 T cell-fate decision. *Immunity*, 51(5):840–855, 2019.

[7] Julia A. Belk, Bence Daniel, and Ansuman T. Satpathy. Epigenetic regulation of T cell exhaustion. *Nature Immunology*, 23(6):848–860, 2022.

[8] Josephine R. Giles et al. Shared and distinct biological circuits in effector, memory and exhausted CD8+ T cells revealed by temporal single-cell transcriptomics and epigenetics. *Nature Immunology*, 23(11):1600–1613, 2022.

[9] Erika L. Pearce et al. Enhancing CD8 T-cell memory by modulating fatty acid metabolism. *Nature*, 460(7251):103–107, 2009.

[10] Padmanee Sharma et al. The next decade of immune checkpoint therapy. *Cancer Discovery*, 11(4):838–857, 2021.

[11] Kexin Huang, Payal Chandak, Qianwen Wang, et al. A foundation model for clinician-centered drug repurposing. *Nature Medicine*, 30(12):3601–3613, 2024. doi: 10.1038/s41591-024-03233-x.

[12] Ayush Noori, Joaquín Polonuer, Katharina Meyer, Bogdan Budnik, Shad Morton, Xinyuan Wang, Sumaiya Nazeen, Yingnan He, Iñaki Arango, Lucas Vittor, Matthew Woodworth, Richard C. Krolewski, Michelle M. Li, Ninning Liu, Tushar Kamath, Evan Macosko, Dylan Ritter, Jalwa Afroz, Alexander B. H. Henderson, Lorenz Studer, Samuel G. Rodriques, Andrew White, Noa Dagan, David A. Clifton, George M. Church, Sudeshna Das, Jenny M. Tam, Vikram Khurana, and Marinka Zitnik. Graph AI generates neurological hypotheses validated in molecular, organoid, and clinical systems. *arXiv preprint arXiv:2512.13724*, 2025.

[13] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, et al. Towards an AI co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

[14] Alexander Novikov et al. AlphaEvolve: A coding agent for scientific and algorithmic discovery. *arXiv preprint arXiv:2506.13131*, 2025.

[15] Daniel S. Himmelstein et al. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *eLife*, 6:e26726, 2017.

[16] Vassilis N. Ioannidis et al. DRKG: A knowledge graph for drug repurposing. *arXiv preprint arXiv:2010.09600*, 2020.

[17] Andres M. Bran, Sam Cox, Oliver Schilter, et al. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6:525–535, 2024. doi: 10.1038/s42256-024-00832-8.

[18] Adithya Kulkarni, Fatimah Alotaibi, Xinyue Zeng, Longfeng Wu, Tong Zeng, Barry Menglong Yao, Minqian Liu, Shuaicheng Zhang, Lifu Huang, and Dawei Zhou. Scientific hypothesis generation and validation: Methods, datasets, and future directions. *arXiv preprint arXiv:2505.04651*, 2025. doi: 10.48550/arXiv.2505.04651.

[19] Payal Chandak, Kexin Huang, and Marinka Zitnik. Building a knowledge graph to enable precision medicine. *Scientific Data*, 10(1):67, 2023. doi: 10.1038/s41597-023-01960-3.

[20] David S. Wishart, Yannick D. Feunang, An Chi Guo, et al. DrugBank 5.0: A major update to the DrugBank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2018. doi: 10.1093/nar/gkx1037.

[21] Damian Szklarczyk, Annika L. Gable, Katerina C. Nastou, et al. The STRING database in 2021: Customizable protein-protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research*, 49(D1):D605–D612, 2021. doi: 10.1093/nar/gkaa1074.

[22] Janet Piñero, Juan Manuel Ramírez-Anguita, Josep Saüch-Pitarch, et al. The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research*, 48(D1):D845–D855, 2020. doi: 10.1093/nar/gkz1021.

[23] Marc Gillespie, Bijay Jassal, Ralf Stephan, et al. The reactome pathway knowledgebase 2022. *Nucleic Acids Research*, 50(D1):D419–D426, 2022. doi: 10.1093/nar/gkab1028.

[24] Michael Ashburner, Catherine A. Ball, Judith A. Blake, et al. Gene Ontology: Tool for the unification of biology. *Nature Genetics*, 25(1):25–29, 2000. doi: 10.1038/75556.

[25] Arthur Liberzon, Aravind Subramanian, Ross Barber, et al. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011. doi: 10.1093/bioinformatics/btr260.

[26] E. John Wherry, Sang-Jun Ha, Susan M. Kaech, et al. Molecular signature of CD8+ T cell exhaustion during chronic viral infection. *Immunity*, 27(4):670–684, 2007. doi: 10.1016/j.immuni.2007.09.006.

[27] Ralf Herbrich, Tom Minka, and Thore Graepel. TrueSkill™: A Bayesian skill rating system. In *Advances in Neural Information Processing Systems*, volume 19, pages 569–576, 2007.

[28] Imre Lakatos. *The Methodology of Scientific Research Programmes: Philosophical Papers Volume 1*. Cambridge University Press, 1978.

[29] Genrich Altshuller. *The Innovation Algorithm: TRIZ, Systematic Innovation and Technical Creativity*. Technical Innovation Center, 1999.

[30] Koichi Araki, Alexandra P. Turner, Virginia Oliva Shaffer, et al. mTOR regulates memory CD8 T-cell differentiation. *Nature*, 460(7251):108–112, 2009. doi: 10.1038/nature08155.

[31] Hazem E. Ghoneim et al. De novo epigenetic programs inhibit PD-1 blockade-mediated T cell rejuvenation. *Cell*, 170(1):142–157, 2017.

[32] Yao Wang et al. Low-dose decitabine priming endows CAR T cells with enhanced and persistent antitumour potential via epigenetic reprogramming. *Nature Communications*, 12(1):409, 2021.

[33] Liang You et al. Decitabine-mediated epigenetic reprograming enhances anti-leukemia efficacy of CD123-targeted chimeric antigen receptor T-cells. *Frontiers in Immunology*, 11:1787, 2020. PMID: 32903685.

[34] Caitlin C. Zebley et al. CD19-CAR T cells undergo exhaustion DNA methylation programming in patients with acute lymphoblastic leukemia. *Science Immunology*, 6(55):eabe4782, 2021. PMC8733956.

[35] Jason A. Dubovsky et al. Ibrutinib is an irreversible molecular inhibitor of ITK driving a Th1-selective pressure in T lymphocytes. *Blood*, 122(15):2539–2549, 2013. PMID: 23886836.

[36] Idit Sagiv-Barfi et al. Therapeutic antitumor immunity by checkpoint blockade is enhanced by ibrutinib, an inhibitor of both BTK and ITK. *Proceedings of the National Academy of Sciences*, 112(9):E966–E972, 2015.

[37] Alice Guyon, Amanda Kussrow, Issa R. Olmsted, Guillaume Sandoz, Darryl J. Bornhop, and Jean-Louis Nahon. Baclofen and other GABAb receptor agents are allosteric modulators of the CXCL12 chemokine receptor CXCR4. *Journal of Neuroscience*, 33(28):11643–11654, 2013. doi: 10.1523/JNEUROSCI.6070-11.2013. GABA-B receptors allosterically modulate CXCR4 chemokine signaling.

[38] Li Chen et al. Hyperuricemia induces T cell dysfunction via NLRP3-mediated inflammasome activation and metabolic reprogramming. *Frontiers in Immunology*, 11:1662, 2020. doi: 10.3389/fimmu.2020.01662. Elevated uric acid dysregulates T-cell function through inflammatory and metabolic mechanisms.

[39] Alireza Mashaghi et al. Neuropeptide substance P and the immune response. *Cellular and Molecular Life Sciences*, 73(22):4249–4264, 2016. doi: 10.1007/s00018-016-2293-z. Substance P modulates T-cell function through NK1 receptor signaling.

AI Co-Scientist Challenge Korea Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state that our hybrid GNN+LLM approach outperforms GNN-only methods (7/7 retained in final pool vs 0/7 in GNN top-100), with Pembrolizumab ranking #1 (validating methodology) and identify novel repurposing candidates (Arbaclofen, Dotinurad, Vofopitant) from non-immunological indications with zero T-cell exhaustion literature. All claims are supported by experimental results in Section 4.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The Discussion section explicitly states limitations: all predictions are computational hypotheses without wet-lab validation, KG may miss emerging targets, and tournament rankings reflect LLM consensus not ground truth efficacy.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This paper presents an empirical framework validated through systematic experiments; formal theoretical proofs are outside the scope of this work as the contributions are methodological and empirical rather than theoretical.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results?

   Answer: [Yes]

   Justification: Section 3 describes the complete methodology including KG construction, GNN training (80/15/5 split, seed=42), agent prompts, and tournament parameters. Supplementary materials provide detailed configuration and resource usage.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code?

   Answer: [No]

   Justification: Our knowledge graph extends PrimeKG and integrates databases (DrugBank, STRING, DisGeNET) that are licensed for research use only and prohibit redistribution. The constructed dataset and pipeline code are therefore available from the corresponding author upon reasonable request for research purposes (see Data Availability section).

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3 specifies GNN training (80/15/5 split, seed=42, 1:1 negative sampling), LLM models (GPT 5.1, Claude Opus 4.5), tournament parameters (3 rounds, 50 max comparisons, TrueSkill rating), and 4 evolution loops. Supplementary materials provide additional configuration details.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: Ablation experiments were conducted with 3 random seeds (42, 43, 44) and we report mean values. With only 3 seeds, standard deviation estimates would be unreliable; instead, we use multiple seeds to verify result stability rather than estimate population variance. The GNN training uses a fixed seed (42) for reproducibility.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Supplementary materials report total duration (2,475.8s), LLM calls (230), tokens (777,481), estimated cost ($4.50), and token usage breakdown by model (GPT 5.1: 691K tokens; Claude Opus 4.5: 86K tokens).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform with the NeurIPS Code of Ethics?

   Answer: [Yes]

   Justification: This research conforms to the NeurIPS Code of Ethics. Drug repurposing candidates are computational hypotheses requiring experimental validation before clinical use.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Positive impact: addresses T-cell exhaustion affecting 40–50% of cancer patients who fail checkpoint inhibitors. Potential negative impact: computational predictions could be misinterpreted as clinical recommendations without proper validation.

11. **Safeguards**

    Question: Does the paper describe safeguards for responsible release of data or models with high misuse risk?

    Answer: [N/A]

    Justification: The paper presents a drug repurposing framework, not a pretrained model or dataset with high misuse risk. All candidates are existing approved drugs requiring standard clinical validation.

12. **Licenses for existing assets**

    Question: Are the creators of assets used in the paper properly credited and are licenses respected?

    Answer: [Yes]

    Justification: PrimeKG is cited and publicly available. All databases (DrugBank, STRING, DisGeNET, Reactome, GO, MSigDB) are properly cited with references.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented?

    Answer: [Yes]

    Justification: The T-cell exhaustion-augmented KG is described in Section 3.1 with construction methodology. Data availability is stated in Methods and Data Availability sections.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include full instructions and compensation details?

    Answer: [N/A]

Justification: This paper does not involve crowdsourcing or research with human subjects.

15. **Institutional Review Board (IRB) Approvals**

Question: Does the paper describe potential risks and IRB approvals for research with human subjects?

Answer: [N/A]

Justification: This paper does not involve research with human subjects.