LILY: A Cancer Gene Prediction Engine Empowered by Biomedical LLMs

Anonymous ACL submission

Abstract

Pinpointing cancer genes (tumor promoters or suppressors) within thousands of cancer-related genes is fundamental to oncogenomics, which studies genetic changes leading to cancer. Approaches to analyzing biological data such as DNA sequence and gene expression for the discovery of cancer-related genes are constrained by their high dimensionality, sparsity, and noise, which impede capturing all relevant connections. Therefore, we propose an alternative and unexplored perspective: Instead of inferring directly from biological data, we systematically integrate existing textual knowledge of gene-cancer associations from the oncogenomics literature to identify genes most strongly involved in cancer-related activities. We introduce **LILY** (Latent, Interaction, Learn, and Yield), a computational hub that bridges and uncovers a substantial volume of promising, novel gene-cancer relationships. It leverages Biomedical Large Language Models (BioLLMs) to extract fragmented information from individual studies and converts these relationships into numerical representations. Then, it interactively refines its knowledge through validation of latent gene-gene and cancer-cancer associations and generates predictions of cancer-related genes with high confidence. Empirical results demonstrate that **LILY** produces highly accurate predictions for cancer-related genes in breast, cervical, lung, prostate, and sarcoma cancers using limited training data. Moreover, its performance incrementally improves as additional data become available, a finding further substantiated by robustness tests and ablation studies.

1 Introduction

Today, of the approximately 20,000 protein-coding genes discovered in the human genome, about 700 have been identified as cancer genes: driver genes with mutations or overexpression that either actively promote tumor progression (known as oncogenes) or suppress it (Martínez-Jiménez et al., 2020; Zhang et al., 2024). For instance, TP53, a tumor suppressor gene, regulates the cellular response to DNA damage and maintains genomic stability through mechanisms such as cell cycle arrest, senescence, and apoptosis (Funk et al., 2025). Inactivation of TP53, or activation of oncogenes like MDM2—which negatively regulates TP53 by promoting its degradation, affects these functions, allowing damaged cells to bypass safeguards and proliferate uncontrollably, leading to tumorigenesis (see Figure 1). Similarly, overexpression of the HER2 gene, common in certain aggressive breast cancers, promotes uncontrolled cell proliferation and survival by activating key signaling pathways such as PI3K/AKT and MAPK. This discovery has led to targeted therapies such as trastuzumab, a monoclonal antibody that specifically inhibits

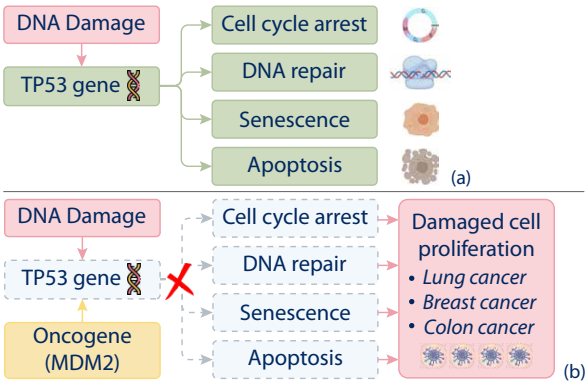


Figure 1: (a) TP53, a tumor suppressor gene, regulates the cell's response to DNA damage through mechanisms like cell cycle arrest, DNA repair, senescence, and apoptosis, helping prevent cancer development. (b) However, when oncogenes are activated or TP53 is inactivated, such as through MDM2, its functions are compromised, allowing damaged cells to proliferate uncontrollably. This promotes tumorigenesis, increasing the risk of cancers of the lung, breast, and colon. Texts with a colored background refer to gene or cancer entities.

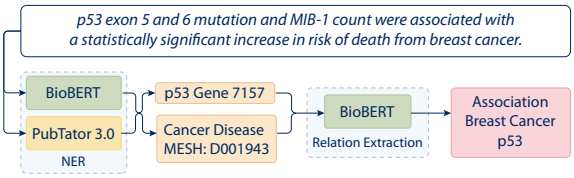


Figure 2: Our pipeline model for NER and RE.

HER2 signaling and improves patient outcomes (Slamon et al., 1987). Accurate targeting of cancer genes enables elucidation of molecular mechanisms, identification of biomarkers for early detection and treatment, and guidance for future research. Although well-known cancer genes such as TP53, MDM2, and HER2 have been established, many additional genes may contribute to oncogenesis and require rigorous experimental validation; however, pinpointing them among nearly 20,000 protein-coding genes for each cancer type remains a formidable challenge. Therefore, identifying candidate genes that are most strongly associated in cancer-related activities is crucial for efficient and effective experimental validation.

Biomedical large language models (BioLLMs) such as BioBERT (Lee et al., 2020) and ClinicalBert (Alsentzer et al., 2019) have excelled in biomedical text mining, patient stratification, and prognostic modeling (Clusmann et al., 2023). It is therefore natural to consider training these BioLLMs on oncogenomics literature from sources including PubMed (NLM, 2025) and OMIM (McKusick, 2007), which offer a rich, high-quality labeled repository of gene–cancer associations derived from clinical studies and expert diagnoses, capturing both experimentally validated associations and observed correlations. However, three challenges remain: (1) most gene–cancer associations are still undiscovered, leaving the training data insufficient despite the literature's richness and validity; (2) the information is inherently fragmented, often from isolated articles (e.g., *"Two genes, called BRCA-1 and BRCA-2, have been identified that appear to be responsible for the majority of familial breast cancer syndromes"* and *"The cancer risks associated with BRCA-2 mutations appear to be somewhat lower than those of BRCA-1" (Mann and Borgen, 1998)*), complicating BioLLMs processing; and (3) LLMs remain susceptible to hallucination, which undermines their ability to accurately identify cancer-related genes for efficient experimental resource allocation (Li et al., 2024b). Therefore, we propose **LILY**, a computational model that leverages BioLLMs for training data collection, integrates such data to model the complex networks underlying gene–cancer associations and produces all predictions simultaneously with high confidence and precision using available information (Cremin et al., 2022; Moon et al., 2023; Hughes et al., 2023; Tian et al., 2024).

We extract gene-cancer dependencies from individual articles in the oncogenomics literature using established BioLLMs and text mining models, including gene–cancer associations, gene-gene interactions (regulatory/co-expression patterns), and cancer–cancer correlations (shared pathways/phenotypic similarities) (Lai et al., 2021a; Kinnersley et al., 2024). These dependencies are converted into standardized numerical representations that capture connection strength and the frequency of repeated mentions, forming three high-dimensional yet sparse matrices that document latent dynamics between gene-cancer, gene-gene, and cancer-cancer as inputs into **LILY**. We developed a novel sparse matrix completion algorithm that interactively optimizes these matrices by leveraging constraints imposed by their interrelationships. The optimized matrices retain biological plausibility (e.g., shared pathways and phenotypic similarities) and yield remarkable performance in predicting novel cancer-related genes with scarce data and substantial improvements as additional data become available (Hoehndorf et al., 2014; Sunde et al., 2024). Our key contributions are:

1. We introduce a novel computational model that integrates oncogenomics literature to predict cancer genes exclusively from BioLLMs-extracted data.
2. We demonstrate that computationally inferring gene–cancer associations, by integrating interactive constraints derived from inferred gene-gene and cancer–cancer relationships, overcomes BioLLMs' limitations in linking fragmented information.
3. We find that incorporating additional interactive constraints among entity relationships may further improve BioLLMs' ability to robustly bridge information beyond gene–cancer associations, such as cancer–symptom and cancer–medicine relationships. Therefore, we provide our collected experimental datasets for future comparative studies.

2 Related Works

2.1 Named Entity Recognition and Relation Extraction in BioLLMs

BioLLMs are tailored to biomedical texts, which

differ significantly from general language (Friedman et al., 2002). Biomedical Named Entity Recognition (NER) identifies domain-specific entities (e.g., genes, cancers, chemicals). To address the resource-intensive, expertise-driven nature of oncogenomics extraction, recent studies have yielded promising results: KECI enhances entity and relation extraction by fusing span graphs with Unified Medical Language System (UMLS) knowledge via collective attention (Lai et al., 2021b). BERT-AMR-KG boosts biomedical information extraction by fusing abstract meaning representation with knowledge graphs via an edge-conditioned graph attention network (Zhang et al., 2021). PubTator 3.0 (Wei et al., 2019) employs AIONER (Luo et al., 2023) for NER and BioREx (Lai et al., 2023) for relation extraction. In our work, we use PubTator 3.0 in a pipeline approach to label oncogenomics articles from sources including OMIM (McKusick, 2007) and PubMed (NLM, 2025), centralizing on cancer–gene relationships. This curated dataset serves as the robust data source for **LILY**.

2.2 Sparse Matrix Completion

Sparse matrix completion has demonstrated significant success in predicting missing values from observed data and inferring unobserved relationships, achieving remarkable results across domains such as recommendation systems, social network analysis, and signal processing (Candes and Recht, 2008; Wen et al., 2012; Bertsimas and Li, 2020; Kim and Chung, 2023; Wang et al., 2023). In **LILY**, after converting textual gene–cancer associations extracted from oncogenomics literature into embeddings, we employ a probabilistic framework for matrix completion to infer missing links in gene-cancer networks, aided by gene-gene interactions and cancer-cancer correlations to substantiate the results. Our model outputs predictions of cancer genes with a predetermined, high level of confidence, while those below the confidence threshold are excluded to preserve the original sparsity of the data (Zhou and Tao, 2011; Li et al., 2024a).

3 The Proposed Method

In this section, we present the detailed theoretical foundations of **LILY**, our proposed model. Key notations are summarized in Table 1.

3.1 Structured Representation of gene-cancer Relationships

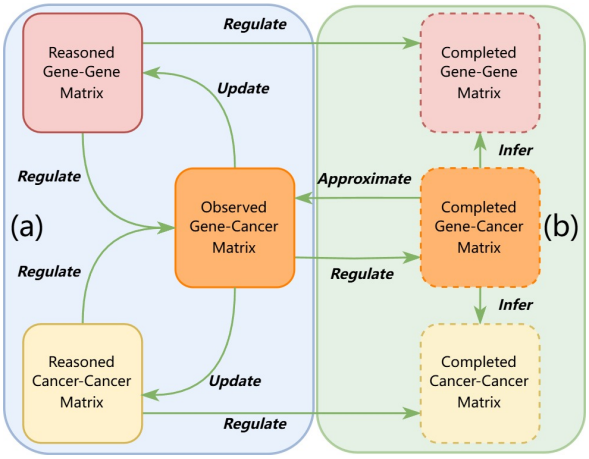


Figure 3: Overview of **LILY**: (a) Interactive updates between the observed gene-cancer matrix and reasoned gene-gene and cancer-cancer matrices constructed by processed oncogenomics data. (b) Completed gene-cancer, gene-gene, and cancer-cancer matrices with high-confidence approximations by solving Eq. 1.

Oncogenetics articles were retrieved from OMIM, PubMed Central, and ClinicalTrials.gov. We developed a pipeline to extract relevant entities and relationships. We performed named entity recognition (NER) using a fine-tuned BioBERT model with a BIO scheme to label each token as beginning (B-), inside (I-), or outside (O) an entity. To enhance coverage, we also employed PubTator 3.0 for extraction. Since PubTator only tags "DISEASE" entities, we additionally extracted specific MESH IDs for different cancer type mentions. PubTator 3.0 has been updated to use AIONER for NER and GNorm2 for gene normalization. While using our own NER module, we allowed partial matching of predicted mentions to fully leverage PubTator 3.0's normalization. We further fine-tuned BioBERT to extract gene-cancer, gene-gene, and cancer-cancer relations (see pipeline in Figure 2), which are converted into numerical representations that form the basis of the relationship matrices used in subsequent computations (see Figure 3).

3.2 The Objective Function

Processed oncogenomics data first forms a gene-cancer matrix $M_{gc} \in \mathbb{R}^{m \times n}$, where m is the number of genes and n is the number of cancer types. Each entry $M_{gc}[i, j]$ quantifies the strength or presence of the association between gene i and cancer type j. However, many entries are missing, posing significant challenges for downstream analysis.

To address this, we develop a sparse matrix completion algorithm to infer missing entries in M_{gc}

Notation	Definition
m	Number of gene types
n	Number of cancer types
i, k	Gene indices
j, l	Cancer type indices
$M_{gc} \in \mathbb{R}^{m \times n}$	Observed gene-cancer matrix
$M_{gg} \in \mathbb{R}^{m \times m}$	Reasoned gene-gene matrix
$M_{cc} \in \mathbb{R}^{n \times n}$	Reasoned cancer-cancer matrix
$M_{gc}^{\text{com}} \in \mathbb{R}^{m \times n}$	Completed gene-cancer matrix
$P_{\Omega_{gc}}(\cdot)$	Observed-entry projection
$P_{\chi^\tau}(\cdot)$	Confidence-mask projection
χ^τ	Binary confidence mask
A_{ij}	Annotated association scores
C_{ij}	Propagated association scores
$f\big(A_{ij}, C_{ij}\big(M_{gc}^{\text{com}}\big)\big)$	Confidence score
τ	Threshold for χ^τ
α	High confidence score for $f(\cdot,\cdot)$
w	Weight for A_{ij}
λ_1	Regularization weight
TopX(i)	Top X related cancers for gene i

Table 1: Notations used in the objective function.

by integrating observed relationships from oncogenomics literature and high-confidence auxiliary information. Specifically, we solve the following convex optimization problem (Kilmer and Martin, 2011; Candès and Recht, 2012; Davis et al., 2021):

$$\begin{aligned}\min_{M_{gc}^{\text{com}}} &\left\|P_{\Omega_{gc}}\left(M_{gc} - M_{gc}^{\text{com}}\right)\right\|_F^2 \\ &+ \lambda_1 \left\|P_{\chi^\tau}\left(M_{gc}^{\text{com}}\right)\right\|_* \\ \text{s.t. } \forall (i,j) \in \text{TopX}(i) &: f\left(A_{ij}, C_{ij}\left(M_{gc}^{\text{com}}\right)\right) \geq \alpha.\end{aligned} \tag{1}$$

Here, $M_{gc}^{\text{com}} \in \mathbb{R}^{m \times n}$ represents the completed gene-cancer matrix. The objective function balances fidelity to the observed data, enforcement of a low-rank structure, alignment with prior knowledge, and statistically robust relationships.

The first term of Eq. 1 preserves the observed entries in the original matrix M_{gc} within M_{gc}^{com}. Specifically, the projection operator $P_{\Omega_{gc}}(\cdot)$ restricts the optimization to the observed entries, which prevents inferred values from overwriting known data and ensures consistency with available observations.

The second term enforces a low-rank structure on M_{gc}^{com}, which facilitates the discovery of fundamental biological patterns and reduces noise. This regularization is applied only to confidence entries, as determined by the projection operator $P_{\chi^\tau}(\cdot)$ and the binary mask $\chi^\tau \in \{0,1\}^{m \times n}$. This mask is generated by applying a threshold τ to the confidence score $f\big(A_{ij}, C_{ij}\big(M_{gc}^{\text{com}}\big)\big)$, which integrates the annotated association score A_{ij} from M_{gc} with the propagated association score C_{ij} from M_{gc}^{com}:

$$\chi^\tau[i,j] = \begin{cases} 1, & \text{if } f\big(A_{ij}, C_{ij}\big(M_{gc}^{\text{com}}\big)\big) \geq \tau \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where

$$f\big(A_{ij}, C_{ij}\left(M_{gc}^{\text{com}}\right)\big) = |w \cdot A_{ij} + C_{ij}|, \tag{3}$$

with w a weight parameter. The use of absolute values in Eq. 3 mitigates errors from the text mining-derived A_{ij} and captures both positive and negative contributions, which enhance robustness. By applying the low-rank constraint, scaled by $\lambda_1 > 0$, exclusively to these high-confidence entries, the model retains flexibility in those entries where the available data do not provide sufficient confidence for reliable prediction.

The constraint of the confidence score τ in Eq. 2 retains gene-cancer associations deemed significant for consideration. A stricter threshold $\alpha > \tau$ in Eq. 1 further ensures that the top X probable cancer types related to gene i, denote as TopX(i) and measured in confidence score in $M_{\text{gc}}^{\text{com}}$, where X is a positive integer, satisfies an even more rigorous criterion:

$$\text{TopX}(i) = \{C_{ij} \in \text{top}_X(\{C_{i1}, \ldots, C_{in}\})\}, \tag{4}$$

where $\text{top}_X(\cdot)$ denotes the X highest values in the set. Specifically, the propagated correlation score C_{ij} is derived from the reasoned gene-gene correlation matrix $M_{gg} \in \mathbb{R}^{m \times m}$ and cancer-cancer correlation matrix $M_{cc} \in \mathbb{R}^{n \times n}$, which encode pairwise relationships based on association patterns in M_{gc}. The propagated correlation score C_{ij} is derived by summing over genes k and cancers l:

$$\begin{aligned} C_{ij} = &\sum_{k=1}^{m} M_{gg}[i,k] \cdot M_{gc}^{\text{com}}[k,j] \\ &+ \sum_{l=1}^{n} M_{cc}[j,l] \cdot M_{gc}^{\text{com}}[i,l], \end{aligned} \tag{5}$$

thus enabling indirect gene-gene relationships to inform the gene-cancer matrix. To construct M_{gg}, we treat each row of M_{gc} as a vector and compute the Pearson correlation coefficient (PCC) between rows i and k (Schober et al., 2018):

$$M_{gg}[i,k] = \text{PCC}(M_{gc}[i,:], M_{gc}[k,:]). \tag{6}$$

Similarly, M_{cc} is built by treating each column of M_{gc} as a vector and computing the PCC between columns j and l:

$$M_{cc}[j,l] = \text{PCC}(M_{gc}[:,j], M_{gc}[:,l]). \tag{7}$$

Collectively, the objective function ensures that M_{gc}^{com} preserves observed data, uncovers underlying biological structure through low-rank constraints, and integrates both direct and propagated information, yielding a robust and interpretable completed gene-cancer matrix.

3.3 Sparse Matrix Completion

To solve the objective function in Eq. 1, we employ the Projected Proximal Method. We decompose the objective function in Eq. 1 into two parts: a smooth component and a non-smooth component. The smooth component is defined as:

$$F(M_{gc}^{\text{com}}) = \|P_{\Omega_{gc}}(M_{gc} - M_{gc}^{\text{com}})\|_F^2, \quad (8)$$

which is differentiable with respect to M_{gc}^{com} and is well-suited to gradient-based optimization. The non-smooth component is given by:

$$R(M_{gc}^{\text{com}}) = \lambda_1 \|P_{\chi^\tau}(M_{gc}^{\text{com}})\|_*. \quad (9)$$

Additionally, we impose the following linear constraints:

$$\mathcal{S} = \Big\{ M_{gc}^{\text{com}} \in \mathbb{R}^{m\times n} \,\Big|\, f\big(A_{ij}, C_{ij}\big(M_{gc}^{\text{com}}\big)\big) \geq \alpha \quad \forall (i,j) \in \text{TopX}(i) \Big\}, \quad (10)$$

where $f\big(A_{ij}, C_{ij}\big(M_{gc}^{\text{com}}\big)\big)$ is linear in M_{gc}^{com}. Consequently, the feasible set $\mathcal{S}$ forms a convex polyhedron–an intersection of half-spaces–which can be efficiently handled with quadratic programming. Due to the convexity of both F (Eq. 8) and R (Eq. 9), the Projected Proximal Method iteratively updates the completed matrix M_{gc}^{com} through gradient descent on F, proximal updates on R, and projection onto $\mathcal{S}$ until convergence.

Gradient Descent Step on $F(M_{gc}^{\text{com}})$: We compute the gradient $\nabla F\big(M_{gc}^{\text{com}}\big)$ of $F(M_{gc}^{\text{com}})$ with respect to M_{gc}^{com} and update the matrix as follows:

$$M_{gc}^{\text{com},(t+\frac{1}{2})} = M_{gc}^{\text{com},(t)} - \eta \cdot \nabla F\big(M_{gc}^{\text{com},(t)}\big), \quad (11)$$

where $\eta > 0$ is the step size. We set $\eta = 10^{-3}$.

Proximal Step on $R(M_{gc}^{\text{com}})$: Given the intermediate matrix $M_{gc}^{\text{com},(t+\frac{1}{2})}$ from Eq. 11, we apply the confidence-mask projection $P_{\chi^\tau}(\cdot)$ as defined in Eq. 2 to retain only high-confidence entries:

$$X' = P_{\chi^\tau}\big(M_{gc}^{\text{com},(t+\frac{1}{2})}\big). \quad (12)$$

We then perform Singular Value Decomposition (SVD) on $X' \in \mathbb{R}^{m\times n}$ and reconstruct the matrix using the thresholded singular values to obtain $\widetilde{X} \in \mathbb{R}^{m\times n}$, the thresholded matrix. The proximal operator, $M_{gc}^{\text{com},(t+\frac{1}{2},\text{svt})}$, is thus expressed as the combination of $\widetilde{X}$ and the entries excluded by $P_{\chi^\tau}(\cdot)$ in the intermediate matrix:

$$M_{gc}^{\text{com},(t+\frac{1}{2},\text{svt})} = \widetilde{X} + (I - P_{\chi^\tau})\big(M_{gc}^{\text{com},(t+\frac{1}{2})}\big), \quad (13)$$

where $(I - P_{\chi^\tau})(\cdot)$ denotes the element-wise complement of the confidence-mask projection.

Projection Step on $\mathcal{S}$: We define the projection step as finding the matrix $Z \in \mathcal{S}$ that minimizes the Frobenius norm distance to $M_{gc}^{\text{com},(t+\frac{1}{2})}$:

$$Z^* = \arg\min_{Z\in\mathcal{S}} \left\| Z - M_{gc}^{\text{com},(t+\frac{1}{2},\text{svt})} \right\|_F^2, \quad (14)$$

where the feasible set $\mathcal{S}$ is defined as in Eq. 10:

$$\mathcal{S} = \Big\{ Z \in \mathbb{R}^{m\times n} \,\Big|\, f\big(A_{ij}, C_{ij}(Z)\big) \geq \alpha, \quad \forall (i,j) \in \text{TopX}(i) \Big\}, \quad (15)$$

with $f\big(A_{ij}, C_{ij}(Z)\big)$ linear in Z. Given the large dimensions m and n are large, we reformulate this projection step in Eq. 14 as a quadratic programming problem. The optimal solution Z^* is then used to update M_{gc}^{com}:

$$M_{gc}^{\text{com},(t+1)} = Z^*. \quad (16)$$

Convergence Check: The Projected Proximal Method iterates through gradient descent, proximal updates, and projection until convergence. With $\varepsilon = 10^{-6}$ the preset tolerance,

$$\|M_{gc}^{\text{com},(t+1)} - M_{gc}^{\text{com},(t)}\|_F < \varepsilon. \quad (17)$$

Computational and Space Complexity: With T iterations until convergence, the total time complexity is $\mathcal{O}(T \cdot (m+n)^2)$. The space complexity is $\mathcal{O}(mn + m^2 + n^2)$.

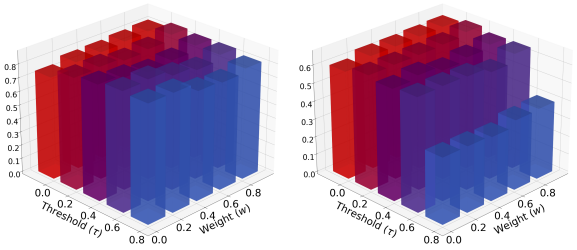
4 Experiments

4.1 Experimental Settings

Datasets: The datasets consist of oncogenomics articles extracted from OMIM and PubMed using a consistent query (e.g., for sarcoma cancer, a rare cancer: "(Sarcoma, Ewing[MESH] AND gene[title/abstract])"). For well-studied cancers (prostate, cervical, breast, and lung), we fixed the number of articles at 10,000 to evaluate our model under limited data conditions, whereas only 1,061

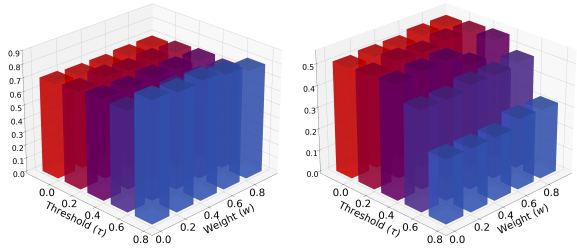
Test Drop (%)	TopX(i), X = 3			TopX(i), X = 5			TopX(i), X = 7			TopX(i), X = 9			TopX(i), X = 11		
	P	**Recall**	**$F_{0.5}$**	**P**	**Recall**	**$F_{0.5}$**	**P**	**Recall**	**$F_{0.5}$**	**P**	**Recall**	**$F_{0.5}$**	**P**	**Recall**	**$F_{0.5}$**
Prostate Cancer Data (10,000 oncogenomics articles, 25,524 relevant lines, 2,965 annotated genes, and 990 cancer types.)															
10%	92.27	47.87	77.83	91.92	45.73	76.47	**92.49**	49.37	78.74	92.17	**53.13**	**80.36**	92.08	46.73	77.11
30%	91.88	45.48	**76.31**	91.46	39.89	72.67	91.91	42.29	74.44	91.49	**45.75**	76.24	**92.00**	42.82	74.81
50%	89.43	33.44	66.99	**90.15**	36.17	69.42	89.13	37.39	69.81	88.97	**39.21**	**70.96**	89.21	37.69	70.06
70%	87.85	31.65	64.83	**88.18**	32.66	65.81	87.07	34.01	66.36	86.99	**36.03**	**67.81**	86.96	33.67	66.50
Cervical Cancer Data (10,000 oncogenomics articles, 17,115 relevant lines, 2,407 annotated genes, and 692 cancer types.)															
10%	**91.95**	42.55	**74.63**	90.91	41.67	73.53	91.36	38.54	71.71	87.91	41.67	71.94	80.93	**46.51**	70.49
30%	**91.03**	39.01	71.86	90.48	33.33	67.37	90.81	41.15	**73.15**	86.75	38.71	69.50	80.36	**42.56**	68.24
50%	**90.32**	32.94	**66.99**	89.47	31.29	65.22	87.88	33.53	66.36	85.51	34.10	65.70	77.31	**40.07**	65.19
70%	**89.29**	30.68	**64.60**	88.31	31.10	64.56	87.93	30.72	64.07	85.25	31.33	63.41	73.00	**32.44**	58.40
Breast Cancer Data (10,000 oncogenomics articles, 38,620 relevant lines, 3,641 annotated genes, and 829 cancer types.)															
10%	**87.54**	42.12	72.01	87.17	**42.55**	**72.06**	84.94	42.29	70.68	84.81	41.88	70.38	84.87	42.08	70.53
30%	**87.22**	41.25	71.32	86.94	41.69	**71.43**	85.17	**41.88**	70.58	83.51	35.84	65.96	83.59	36.06	66.15
50%	**87.33**	30.05	63.22	85.80	**34.63**	**66.23**	83.42	34.51	65.00	82.93	32.54	63.31	82.64	33.01	63.54
70%	**86.08**	27.30	60.17	85.25	27.23	59.77	83.69	30.10	61.72	83.56	31.12	62.50	83.55	**32.40**	**63.50**
Lung Cancer Data (10,000 oncogenomics articles, 60,532 relevant lines, 6,242 annotated genes, and 1,716 cancer types.)															
10%	**86.21**	40.32	70.23	84.85	**45.16**	**72.16**	84.38	43.55	71.05	83.87	41.94	69.89	84.00	36.21	66.46
30%	**85.71**	38.76	69.00	84.38	**43.55**	**71.05**	83.33	40.32	68.68	82.76	38.71	67.41	83.03	35.71	65.64
50%	**85.16**	31.58	63.59	81.82	32.77	62.97	82.61	33.33	63.76	83.36	**35.01**	**65.37**	81.92	32.14	62.55
70%	**85.00**	28.33	60.71	80.95	30.00	60.43	80.00	30.53	60.41	77.27	28.33	57.43	82.61	**31.67**	**62.50**
Sarcoma Cancer Data (1,061 oncogenomics literature articles, 5,679 relevant lines, 679 annotated genes, and 283 cancer types.)															
10%	80.00	41.03	67.23	80.95	43.59	69.11	**81.40**	44.28	69.71	80.65	44.87	69.55	78.57	**52.38**	**71.43**
30%	**82.76**	34.29	64.52	80.65	35.71	64.43	80.00	**42.11**	**67.80**	81.82	38.57	66.83	73.81	41.33	63.79
50%	72.22	22.41	50.00	73.68	24.14	52.24	**75.00**	25.86	54.35	71.43	28.28	54.73	71.05	**38.03**	**60.54**
70%	66.67	20.34	45.80	68.42	22.03	48.15	70.00	23.73	50.36	**71.04**	25.42	52.28	67.63	**34.85**	**56.93**

Table 2: Performance of **LILY** in predicting cancer-related genes for prostate, cervical, breast, lung, and sarcoma cancers under varying data availability, controlled by **Test Drop (%)**, and varying influence of available data, controlled by **TopX**(i), evaluated by Precision (P), Recall, and $F_{0.5}$-score. For each cancer type and experimental condition (i.e., **Test Drop (%)** and **TopX**(i)), the best and second-best results are bolded and underlined, respectively.



(a) Precision (b) $F_{0.5}$-score

Figure 4: Lung cancer-related gene predictions by **LILY**: Precision and $F_{0.5}$-score for various τ and w values, with 50% data dropped and X = 7 in TopX(i).



(a) Precision (b) $F_{0.5}$-score

Figure 5: Sarcoma cancer-related gene predictions by **LILY**: Precision and $F_{0.5}$-score for various τ and w values, with 50% data dropped and X = 7 in TopX(i).

articles were available for sarcoma cancer due to its rarity. Data quality was ensured via preprocessing and filtering, employing named entity recognition (NER) to extract relevant entities and relation extraction (RE) to classify relationships as *association*, *positive correlation*, or *negative correlation*. Relations are then categorized into three types, cancer–gene, gene–gene, and cancer–cancer.

Each relationship was scored (1 for *association*, 2 for *positive correlation*, and -1 for *negative correlation*) and aggregated as a weighted average across identical relationships to enhance robustness. These scores, quantifying the strength of each association, serve as inputs to our model for predicting novel cancer-related genes.

Hyperparameters and Evaluation Metrics: Unless otherwise noted, the hyperparameters in Eq. 1 are set as follows: $w = 0.2$, $\tau = 0.2$, $\alpha = 0.8$, and $\lambda_1 = 0.1$, values at which performance peaks. w, τ, and α are tunable within the interval $[0, 1]$. The

Models	Test Drop 10% Data			Test Drop 30% Data			Test Drop 50% Data			Test Drop 70% Data		
	P	Recall	$F_{0.5}$	P	Recall	$F_{0.5}$	P	Recall	$F_{0.5}$	P	Recall	$F_{0.5}$
Prostate Cancer Data & TopX$(i), X = 5$.												
LILY Baseline[1]	61.21	**82.69**	64.56	54.11	**70.23**	56.72	46.28	**62.14**	48.77	13.33	**60.00**	15.79
LILY Baseline[2]	80.27	46.98	70.31	77.17	40.46	65.32	72.92	33.55	59.06	69.23	28.62	53.93
LILY	**91.92**	45.73	**76.47**	**91.46**	39.89	**72.67**	**90.15**	36.17	**69.42**	**88.18**	32.66	**65.81**
Breast Cancer Data & TopX$(i), X = 5$.												
LILY Baseline[1]	60.45	**81.03**	63.69	56.55	**72.95**	59.22	39.55	**54.04**	41.79	14.71	**71.43**	17.48
LILY Baseline[2]	77.52	40.40	65.49	77.14	35.75	62.63	63.16	25.47	48.74	57.90	22.62	44.13
LILY	**87.17**	42.55	**72.06**	**86.94**	41.69	**71.43**	**85.80**	34.63	**66.23**	**85.25**	27.23	**59.77**
Sarcoma Cancer Data & TopX$(i), X = 5$.												
LILY Baseline[1]	57.53	**87.50**	61.76	50.79	**72.73**	54.05	48.08	**71.43**	51.44	9.68	**42.86**	11.45
LILY Baseline[2]	74.58	51.77	68.54	68.63	43.21	61.40	68.18	41.67	**60.48**	65.00	38.81	**57.27**
LILY	**80.95**	43.59	**69.11**	**80.65**	35.71	**64.43**	**73.68**	24.14	52.24	**68.42**	22.03	48.15

Table 3: Performance of **LILY Baseline**[1], **LILY Baseline**[2], and **LILY** in predicting cancer-related genes for prostate, breast, and sarcoma cancers under varying data availability with fixed influence of available data, evaluated by Precision (P), Recall, and $F_{0.5}$-score. The best results are bolded, and the second-best are underlined.

parameter X in TopX(i), which regulates the number of top-related cancers per gene, modulates the influence of available data; higher X corresponds to greater influence. Performance is assessed using precision (P), recall, and the F_β-score, with $\beta = 0.5$ to prioritize precision over recall due to our goal of identifying the most probable cancer-related genes among numerous potential linkages for efficient experimental resource allocation.

Relevant Models: Our work is inspired by prior efforts leveraging BioLLMs to extract association information among genes and diseases, including **DISEASES** (Pletscher-Frankild et al., 2015), **GeneSemantics** (Miller et al., 2022), **GatorTron** (Yang et al., 2022a), **MSK-CHORD** (Jee et al., 2024), and **Teacher-Student** Framework (Kehl et al., 2024). However, no previous study has attempted a BioLLMs-enabled approach to predict cancer genes by integrating fragmented information. Therefore, we propose two baseline methods: **LILY Baseline**[1] adopts the computational framework of **LILY** without confidence-score threshold for predictions. **LILY Baseline**[2] uses only the gene-cancer associations extracted by BioLLMs on the same computational framework and omits gene-gene and cancer-cancer associations.

4.2 Experimental Results

Table 2 presents the prediction results of cancer-related genes by **LILY**. We assess its performance under varying data availability by dropping 10%, 30%, 50%, and 70% of the original dataset (Test Drop %) and adjusting X in TopX(i) to modulate data influence, verifying the predictions against the ground-truth. The results demonstrate:

1. Under 10% data drop, **LILY** yields high precision on well-studied cancers: prostate (92.49%), cervical (91.95%), breast (87.54%), and lung (86.21%). The under-researched sarcoma cancer achieves 81.40%. Under 70% data drop, fixing $X = 7$ in TopX(i), the precision declines by 5.86% (prostate), 3.75% (cervical), 1.47% (breast), 5.19% (lung), and 14.00% (sarcoma), indicating that limited data affects less-studied cancers more severely.
2. Recall decreases with less data availability but is partially offset if available data exerts greater influence. In breast cancer, recall declines by 32.74% with $X = 3$ and by 23.00% with $X = 11$ as the data drop increases from 10% to 70%; in sarcoma cancer, recall declines by 50.43% with $X = 3$ and by 33.47% with $X = 11$. It is suggested that increased data availability enhances the detection of true cancer-related genes while amplifying the impact of available data can mitigate recall reduction.
3. For prostate, cervical, breast, and lung cancers, the F_β-score remains between 60% and 80% with minimal variance across different X settings in TopX(i) at a fixed data drop. In contrast, predictions on sarcoma cancer exhibits substantial variability, with F_β-score ranging from 50.00% to 60.54% at a 50% data drop and from 45.80% to 56.93% at a 70% drop, indicating that limited data impairs the balance between precision and recall.

Table 3 compares the performance of **LILY** with the two baseline models. **LILY** consistently achieves the highest precision, while **LILY Base-**

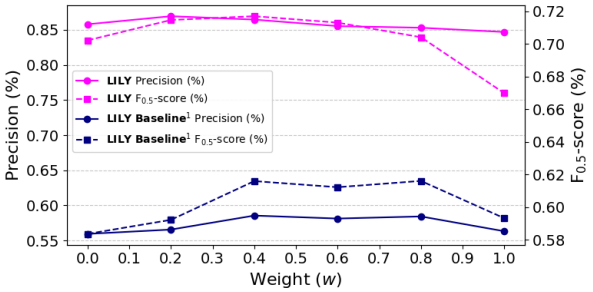


Figure 6: Prediction of breast cancer-related genes by **LILY** and **LILY Baseline**[1] evaluated under various w, with 30% data drop, and X $= 5$ in TopX(i).

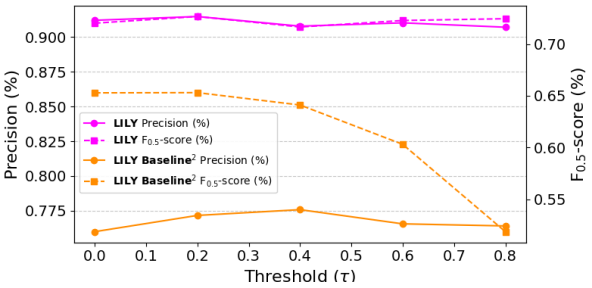


Figure 7: Prediction of prostate cancer-related genes by **LILY** and **LILY Baseline**[2] evaluated under various τ, 30% data drop, and X $= 5$ in TopX(i).

Gene	Score	Relation	Information Source
PDK1	1.013	✓	(Peng et al., 2018)
RBBP8	0.969	✓	(Zarrizi et al., 2020)
BCL3	0.965	✓	(Turnham et al., 2024)
STC2	0.868	✓	(Qie et al., 2024)
TFF1	0.868	?	(Buache et al., 2011)
TFF3	0.868	✓	(Yang et al., 2022b)
MDM2	0.862	✓	(Wang et al., 2014)
RAD54L	0.855	✓	(Gonzalez et al., 1999)
MIR23AHG	0.855	?	(Entezari et al., 2024)
ATF1	0.855	✓	(Huang et al., 2016)
MTND6P4	0.855	?	(Pangeni et al., 2022)
MIR3193	0.840	?	Not Found.
NCAN	0.840	?	(Williams et al., 2024)
TBX5	0.840	✓	(Network, 2012)
CTHRC1	0.840	✓	(Lee et al., 2016)

Table 4: Prediction of novel breast cancer-related genes with data extracted from 10,000 oncogenomics articles.

line[1] shows the lowest precision and F_β-score yet the highest recall. In contrast, **LILY Baseline**[2] attains slightly lower precision but higher recall than **LILY**, resulting in a superior F_β-score on resource-scarce sarcoma data under low data availability. These results indicate: (1) This computational framework covers a broad range of potential gene candidates, but applying a high-confidence threshold is necessary for reliable predictions. (2) Directly using gene-cancer associations from BioLLMs is effective; however, incorporating computed gene-gene and cancer-cancer correlations bridges fragmented information and significantly enhances performance. (3) As data availability decreases (from a 10% to a 70% drop), both baseline models exhibit dramatic performance declines, whereas **LILY** experiences only a mild decrease (3.74% in prostate, 1.92% in breast, and 12.53% in sarcoma). This suggests that even with incomplete direct gene-cancer data, reasoned gene-cancer relationships help sustain the model's performance. **Robustness Analysis and Ablation Studies**: Figures 4 and 5 show that **LILY** demonstrates stable performance across various w and τ combinations, except at $\tau = 0.8$, where precision and $F_{0.5}$-score fluctuate due to an overly high confidence threshold. Figures 6 and 7 further demonstrate that **LILY** consistently outperforms **LILY Baseline**[1] and **LILY Baseline**[2], confirming that both the confidence-score threshold and the reasoning component for gene-gene and cancer–cancer associations are indispensable. Notably, **LILY** achieves peak precision and $F_{0.5}$-score at $\tau = 0.2$ and $w = 0.2$, which we adopt as the optimal settings of parameters.

4.3 Novel Predictions

Table 4 lists the top 15 predicted breast cancer-related genes ranked by confidence score. Trained only on data collected from 10,000 oncogenomics articles, **LILY** identifies novel cancer-related genes, some experimentally validated and others only peripherally noted, and covers protein-coding (e.g., TFF1) and even non-coding genes (e.g., MIR23AHG). Since the training data comprise only a small fraction of potential gene–cancer associations, **LILY**'s accurate inference with limited data demonstrates its efficacy and suggests that incorporating more data and expanding the gene-cancer database will further enhance performance.

5 Conclusion

In this paper, we propose a novel computational model empowered by BioLLMs for integrating gene-cancer networks and predicting novel relations. Trained exclusively on data processed from oncogenomics literature, the model generates highly precise predictions even with limited data and demonstrates the potential for enhanced performance through scalability to larger datasets. It underscores the need for enhanced collaboration with biomedical labs and offers new insights into addressing limitations in current BioLLMs.

Limitations

One limitation arises from the data collection process. To ensure reproducibility and optimize model robustness, we standardize data collection from oncogenomics articles extracted from PubMed and OMIM using a consistent query (e.g., for sarcoma cancer: "(Sarcoma, Ewing[MESH] AND gene[title/abstract])"). This query selects relevant, up-to-date oncogenomics articles, making the data susceptible to bias due to temporal shifts in research focus and search engine dynamics. Another limitation is that, although our model predicts highly probable cancer-related genes, these predictions serve solely as suggestions for rigorous biomedical laboratory testing rather than conclusive identifications. Finally, the model is currently limited to predicting genes for one cancer type at a time, requiring separate data extraction and training for each cancer type, as it does not yet support simultaneous multi-cancer predictions.

Ethics Statement

Our method for extracting gene-cancer association data from oncogenomics literature, sourced from PubMed and OMIM and processed using BioLLMs, adheres to the ethical framework established by the National Library of Medicine (NLM) and the National Center for Biotechnology Information (NCBI). The disclaimers emphasize that these platforms function as aggregators of scientific research rather than publishers and do not provide direct medical advice or endorsements. By using the data strictly for research purposes and not for clinical decision-making or commercial advertising, we strictly follow the stipulation that users should consult qualified healthcare professionals for personal medical issues. Furthermore, we acknowledge the importance of upholding copyright and intellectual property rights in accordance with NCBI's policies. We ensure that all data usage complies with fair use and legal guidelines while providing appropriate attribution to the data providers. Throughout our research, we adhere to rigorous scientific standards, maintain transparency, and responsibly manage potentially sensitive oncogenomics information in accordance with the ethical guidelines outlined by the NLM and NCBI.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78. Association for Computational Linguistics.

Heinz H. Bauschke and Patrick L. Combettes. 2017. *Correction to: Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer.

Amir Beck and Marc Teboulle. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

Dimitris Bertsimas and Michael Lingzhi Li. 2020. Fast exact matrix completion: A unified optimization framework for matrix completion. *Journal of Machine Learning Research*, 21:1–43.

E. Buache, N. Etique, F. Alpy, I. Stoll, M. Muckensturm, B. Reina-San-Martin, M. P. Chenard, C. Tomasetto, and M. C. Rio. 2011. Deficiency in trefoil factor 1 (tff1) increases tumorigenicity of human breast cancer cells and mammary tumor development in tff1-knockout mice. *Oncogene*, 30(29):3261–3273.

Emmanuel Candès and Benjamin Recht. 2012. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119.

Emmanuel J. Candes and Benjamin Recht. 2008. Exact low-rank matrix completion via convex optimization. In *46th Annual Allerton Conference on Communication, Control, and Computing*, pages 806–812.

Jan Clusmann, Fiona R. Kolbinger, Hannah Sophie Muti, Zunamys I. Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P. Veldhuizen, et al. 2023. The future landscape of large language models in medicine. *Communications Medicine*, 3(1):141.

Patrick L. Combettes and Jean-Christophe Pesquet. 2011. Proximal splitting methods in signal processing. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212.

Conor John Cremin, Sabyasachi Dash, and Xiaofeng Huang. 2022. Big data: Historic advances and emerging trends in biomedical research. *Current Research in Biotechnology*, 4:138–151.

Damek Davis, Dmitriy Drusvyatskiy, Lin Xiao, and Junyu Zhang. 2021. From low probability to high confidence in stochastic convex optimization. *Journal of Machine Learning Research*, 22(49):1–38.

Maryam Entezari, Bahram M. Soltani, and Majid Sadeghizadeh. 2024. Microrna-203a inhibits breast cancer progression through the pi3k/akt and wnt pathways. *Scientific Reports*, 14(1):4715.

Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *Journal of Biomedical Informatics*, 35(4):222–235.

Julianne S. Funk, Maria Klimovich, Daniel Drangenstein, Ole Pielhoop, Pascal Hunold, Anna Borowek, Maxim Noeparast, Evangelos Pavlakis, Michelle Neumann, Dimitrios-Ilias Balourdas, Katharina Kochhan, Nastasja Merle, Imke Bullwinkel, Michael Wanzel, Sabrina Elmshäuser, Julia Teply-Szymanski, Andrea Nist, Tara Procida, Marek Bartkuhn, Katharina Humpert, Marco Mernberger, Rajkumar Savai, Thierry Soussi, Andreas C. Joerger, and Thorsten Stiewe. 2025. Deep crispr mutagenesis characterizes the functional diversity of tp53 mutations. *Nature Genetics*, 57(1):140–153.

R. Gonzalez, J. M. Silva, G. Dominguez, J. M. Garcia, G. Martinez, J. Vargas, M. Provencio, P. Espana, and F. Bonilla. 1999. Detection of loss of heterozygosity at rad51, rad52, rad54 and brca1 and brca2 loci in breast cancer: Pathological correlations. *British Journal of Cancer*, 81(3):503–509.

Robert Hoehndorf, Paul N. Schofield, and Georgios V. Gkoutos. 2014. Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases. *Scientific Reports*, 5.

Guo-Liang Huang, Dan Liao, Hua Chen, Yan Lu, Liyong Chen, Huahui Li, Binbin Li, Weilong Liu, Caiguo Ye, Tong Li, et al. 2016. The protein level and transcription activity of activating transcription factor 1 is regulated by prolyl isomerase pin1 in nasopharyngeal carcinoma progression. *Cell Death & Disease*, 7(12):e2571.

Laura D. Hughes, Ginger Tsueng, Jack DiGiovanna, Thomas D. Horvath, Luke V. Rasmussen, Tor C. Savidge, Thomas Stoeger, Serdar Turkarslan, Qinglong Wu, Chun Ting Wu, Andrew I. Su, and Lars Pache. 2023. Addressing barriers in fair data practices for biomedical data. *Scientific Data*, 10.

Justin Jee, Christopher Fong, Karl Pichotta, Anisha Luthra, Michele Waters, Chenlian Fu, Mirella Altoe, Si-Yang Liu, Steven Maron, Mehnaj Ahmed, Susie Kim, Mono Pirun, Walid Chatila, Ino Bruijn, Arfath Pasha, Ritika Kundra, Benjamin Gross, Brooke Mastrogiacomo, and Nikolaus Schultz. 2024. Automated real-world data integration improves cancer outcome prediction. *Nature*.

Kenneth L. Kehl, Justin Jee, Karl Pichotta, Morgan A. Paul, Pavel Trukhanov, Christopher J. Fong, Michele Waters, Ziad Bakouny, Wenxin Xu, Toni K. Choueiri, Chelsea Nichols, Deborah Schrag, and Nikolaus Schultz. 2024. Shareable artificial intelligence to extract cancer outcomes from electronic health records for precision oncology research. *Nature Communications*, 15.

Misha E. Kilmer and Carla D. Martin. 2011. Factorization strategies for third-order tensors. *Linear Algebra and its Applications*, 435(3):641–658. Special Issue: Dedication to Pete Stewart on the Occasion of His 70th Birthday.

Daesung Kim and Hye Won Chung. 2023. Rank-1 matrix completion with gradient descent and small random initialization. In *Advances in Neural Information Processing Systems*, volume 36, pages 10530–10566.

Ben Kinnersley, Amit Sud, Andrew Everall, Alex J. Cornish, Daniel Chubb, Richard Culliford, Andreas J. Gruber, Adrian Lärkeryd, Costas Mitsopoulos, David Wedge, and Richard Houlston. 2024. Analysis of 10,478 cancer genomes identifies candidate driver genes and opportunities for precision oncology. *Nature Genetics*, 56(9):1868–1877.

Po-Ting Lai, Chih-Hsuan Wei, Ling Luo, Qingyu Chen, and Zhiyong Lu. 2023. Biorex: improving biomedical relation extraction by leveraging heterogeneous datasets. *Journal of Biomedical Informatics*, 146:104487.

Tuan Lai, Heng Ji, ChengXiang Zhai, and Quan Hung Tran. 2021a. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6248–6260. Association for Computational Linguistics.

Tuan Lai, Heng Ji, ChengXiang Zhai, and Quan Hung Tran. 2021b. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6248–6260. Association for Computational Linguistics.

Jaemin Lee, Jinhoi Song, Eun-Soo Kwon, Seongyea Jo, Min Kyung Kang, Yeon Jeong Kim, Yeonsil Hwang, Hosung Bae, Tae Heung Kang, Suhwan Chang, et al. 2016. Cthrc1 promotes angiogenesis by recruiting tie2-expressing monocytes to pancreatic tumors. *Experimental & Molecular Medicine*, 48(9):e261.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Jiangyuan Li, Jiayi Wang, Raymond K. W. Wong, and Kwun Chuen Gary Chan. 2024a. A pairwise pseudo-likelihood approach for matrix completion with informative missingness. In *Advances in Neural Information Processing Systems*, volume 37, pages 10735–10769. Curran Associates, Inc.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024b. The

dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899. Association for Computational Linguistics.

Ling Luo, Chih-Hsuan Wei, Po-Ting Lai, Robert Leaman, Qingyu Chen, and Zhiyong Lu. 2023. Aioner: all-in-one scheme-based biomedical named entity recognition using deep learning. *Bioinformatics*, 39(5):btad310.

G. Bruce Mann and Patrick I. Borgen. 1998. Breast cancer genes and the surgeon. *Journal of Surgical Oncology*, 67(4):267–274.

Francisco Martínez-Jiménez, Ferran Muiños, Inés Sentís, Jordi Deu-Pons, Iker Reyes-Salazar, Claudia Arnedo-Pac, Loris Mularoni, Oriol Pich, Jose Bonet, Hanna Kranas, Abel González-Pérez, and Núria López-Bigas. 2020. A compendium of mutational cancer driver genes. *Nature Reviews Cancer*, 20:555–572.

Victor A McKusick. 2007. Mendelian inheritance in man and its online version, omim. *The American Journal of Human Genetics*, 80(4):588–604.

Danielle Miller, Adi Stern, and David Burstein. 2022. Deciphering microbial gene function using natural language processing. *Nature Communications*, 13(1).

Intae Moon, Jaclyn LoPiccolo, Sylvan Baca, Lynette Sholl, Kenneth Kehl, Michael Hassett, David Liu, Deborah Schrag, and Alexander Gusev. 2023. Machine learning for genetics-based classification and treatment response prediction in cancer of unknown primary. *Nature Medicine*, 29:1–11.

The Cancer Genome Atlas Network. 2012. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70.

National Library of Medicine NLM. 2025. Pubmed. `https://pubmed.ncbi.nlm.nih.gov/`. Accessed: 2025-01-15.

Rajendra P. Pangeni, Ivonne Olivaries, David Huen, Vannessa C. Buzatto, Timothy P. Dawson, Katherine M. Ashton, Charles Davis, Andrew R. Brodbelt, Michael D. Jenkinson, Ivan Bièche, et al. 2022. Genome-wide methylation analyses identifies noncoding rna genes dysregulated in breast tumours that metastasise to the brain. *Scientific Reports*, 12(1):1102.

F. Peng, J. H. Wang, W. J. Fan, Y. T. Meng, M. M. Li, T. T. Li, B. Cui, H. F. Wang, Y. Zhao, F. An, et al. 2018. Glycolysis gatekeeper pdk1 reprograms breast cancer stem cells under hypoxia. *Oncogene*, 37(8):1062–1074.

Sune Pletscher-Frankild, Albert Pallejà, Kalliopi Tsafou, Janos X. Binder, and Lars Juhl Jensen. 2015. Diseases: Text mining and data integration of disease–gene associations. *Methods*, 74:83–89.

Shuo Qie, Haijuan Xiong, Yaqi Liu, Chenhui Yan, Yalei Wang, Lifeng Tian, Chenguang Wang, and Nianli Sang. 2024. Stanniocalcin 2 governs cancer cell adaptation to nutrient insufficiency through alleviation of oxidative stress. *Cell Death & Disease*, 15(8):567.

Patrick Schober, Christa Boer, and Lothar A. Schwarte. 2018. Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*, 126:1763–1768.

Dennis J Slamon, Gary M Clark, Steven G Wong, Wendy J Levin, Axel Ullrich, and William L McGuire. 1987. Human breast cancer: correlation of relapse and survival with amplification of the her-2/neu oncogene. *Science*, 235(4785):177–182.

Hans Fredrik Sunde, Nikolai Haahjem Eftedal, Rosa Cheesman, Elizabeth C. Corfield, Thomas H. Kleppesto, Anne Caroline Seierstad, Eivind Ystrom, Espen Moen Eilertsen, and Fartein Ask Torvik. 2024. Genetic similarity between relatives provides evidence on the presence and history of assortative mating. *Nature Communications*.

Shubo Tian, Qiao Jin, Lana Yeganova, Po-Ting Lai, Qingqing Zhu, Xiuying Chen, Yifan Yang, Qingyu Chen, Won Kim, Donald C. Comeau, Rezarta Islamaj, Aadit Kapoor, Xin Gao, and Zhiyong Lu. 2024. Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, 25(1):bbad493.

Daniel J. Turnham, Hannah Smith, and Richard W. E. Clarkson. 2024. Suppression of bcl3 disrupts viability of breast cancer cells through both p53-dependent and p53-independent mechanisms via loss of nf-b signalling. *Biomedicines*, 12(1):143.

Lingxiao Wang, Boxin Zhao, and Mladen Kolar. 2023. Differentially private matrix completion through low-rank matrix factorization. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 5731–5748.

Wei Wang, Jiang-Jiang Qin, Sukesh Voruganti, Kalkunte S. Srivenugopal, Subhasree Nag, Shivaputra Patil, Horrick Sharma, Ming-Hai Wang, Hui Wang, John K. Buolamwini, et al. 2014. The pyrido [b] indole mdm2 inhibitor sp-141 exerts potent therapeutic effects in breast cancer models. *Nature Communications*, 5(1):5086.

Chih-Hsuan Wei, Alexis Allot, Robert Leaman, and Zhiyong Lu. 2019. Pubtator central: automated concept annotation for biomedical full text articles. *Nucleic acids research*, 47(W1):W587–W593.

Zaiwen Wen, Wotao Yin, and Yin Zhang. 2012. Solving a low-rank factorization model for matrix completion

by a nonlinear successive over-relaxation algorithm. *Mathematical Programming Computation*, 4.

Marc J. Williams, Michael U. J. Oliphant, Vinci Au, Cathy Liu, Caroline Baril, Ciara O'Flanagan, Daniel Lai, Sean Beatty, Michael Van Vliet, Jacky C. H. Yiu, et al. 2024. Luminal breast epithelial cells of brca1 or brca2 mutation carriers and noncarriers harbor common breast cancer copy number alterations. *Nature Genetics*, pages 1–10.

Xi Yang, Aokun Chen, Nima M. Pournejatian, Hoo-Chang Shin, Kaleb E. Smith, Christopher Parisien, Colin B. Compas, Cheryl Martin, Anthony B. Costa, Mona G. Flores, Ying Zhang, Tanja Magoc, Christopher A. Harle, Gloria P. Lipori, Duane A. Mitchell, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, and Yonghui Wu. 2022a. A large language model for electronic health records. *NPJ Digital Medicine*, 5.

Yiqi Yang, Ziyang Lin, Quanyou Lin, Weijian Bei, and Jiao Guo. 2022b. Pathological and therapeutic roles of bioactive peptide trefoil factor 3 in diverse diseases: Recent progress and perspective. *Cell Death & Disease*, 13(1):62.

Reihaneh Zarrizi, Martin R. Higgs, Karolin Voßgröne, Maria Rossing, Birgitte Bertelsen, Muthiah Bose, Arne Nedergaard Kousholt, Heike Rösner, Bent Ejlertsen, Grant S. Stewart, et al. 2020. Germline rbbp8 variants associated with early-onset breast cancer compromise replication fork stability. *The Journal of Clinical Investigation*, 130(8):4069–4080.

Shaosen Zhang, Xinyi Xiao, Yonglin Yi, Xinyu Wang, Lingxuan Zhu, Yanrong Shen, Dongxin Lin, and Chen Wu. 2024. Tumor initiation and early tumorigenesis: molecular mechanisms and interventional targets. *Signal Transduction and Targeted Therapy*, 9:149.

Zixuan Zhang, Nikolaus Parulian, Heng Ji, Ahmed Elsayed, Skatje Myers, and Martha Palmer. 2021. Fine-grained information extraction from biomedical literature based on knowledge-enriched Abstract Meaning Representation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6261–6270. Association for Computational Linguistics.

Tianyi Zhou and Dacheng Tao. 2011. Godec: Randomized low-rank and sparse matrix decomposition in noisy case. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ICML.

A Computational and Space Complexity

The overall computational complexity of **LILY** is primarily determined by the sparse matrix completion algorithm via the projected proximal method. The gradient descent step requires $\mathcal{O}(mn)$ for matrix operations, while the proximal step performs SVD in $\mathcal{O}(\min(mn^2, nm^2))$ time. The projection step, reformulated as a sparse convex quadratic programming problem, requires $\mathcal{O}((m+n)^2)$ time per iteration and dominates the cost. With T iterations until convergence, the total time complexity is $\mathcal{O}(T \cdot (m+n)^2)$. The space complexity is $\mathcal{O}(mn + m^2 + n^2)$, accounting for the observed gene-cancer matrix M_{gc} and the reasoned gene-gene correlation matrix M_{gg} and cancer-cancer correlation matrix M_{cc}.

B Quadratic Programming Solution to Projection Step

We need to enforce, for each TopX gene-cancer pair (i, j) in Eq. 1,

$$\big|\, w \cdot A_{ij} + C_{ij}(Z) \big| \;\geq\; \alpha, \tag{18}$$

where $w \cdot A_{ij}$ is the annotated text-mined score, and

$$\begin{aligned} C_{ij}(Z) = &\sum_{k=1}^{m} M_{gg}[i,k] \cdot M_{gc}^{\text{com}}[k,j] \\ &+ \sum_{l=1}^{n} M_{cc}[j,l] \cdot M_{gc}^{\text{com}}[i,l]. \end{aligned} \tag{19}$$

To remove the absolute value, introduce an auxiliary variable $s_{ij} \geq 0$. Then Eq. 18 becomes:

$$\begin{aligned} s_{ij} \;&\geq\; w \cdot A_{ij} + \sum_{k} M_{gg}[i,k]\, Z[k,j] \\ &\qquad + \sum_{l} M_{cc}[j,l]\, Z[i,l], \\ s_{ij} \;&\geq\; -\Big(w \cdot A_{ij} + \sum_{k} M_{gg}[i,k]\, Z[k,j] \\ &\qquad + \sum_{l} M_{cc}[j,l]\, Z[i,l] \Big), \\ s_{ij} \;&\geq\; \alpha, \quad s_{ij} \;\geq\; 0. \end{aligned} \tag{20}$$

These inequalities ensure $|\, w \cdot A_{ij} + C_{ij}(Z)| \leq s_{ij}$ and $s_{ij} \geq \alpha$; thus, $|\, w \cdot A_{ij} + C_{ij}(Z)| \geq \alpha$.

After the gradient and proximal updates, let $\widetilde{X} = M_{gc}^{\text{com},(t+\frac{1}{2},\text{svt})}$. The next iterate Z is found by solving:

$$\begin{aligned} \min_{Z,\,\{s_{ij}\}} \quad & \sum_{p,q} \big(Z[p,q] - \widetilde{X}[p,q]\big)^2 \\ \text{s.t.} \quad & \text{Inequalities in Eq. 20}, \quad \forall\, (i,j) \in \text{TopX}(i). \end{aligned} \tag{21}$$

Since $\|Z - \widetilde{X}\|_F^2$ is a standard least-squares objective, and Eq. 20 is linear, Eq. 21 is a standard Quadratic Program (QP) suitable for widely available solvers. The solution Z^* exactly satisfies Eq. 18 and is thus used to update $M_{gc}^{\text{com},(t+1)}$.

C Proof of Guaranteed Convergence of the Objection Function

As addressed in Section 3.3 Sparse Matrix Completion, we decompose the objective into a smooth component $F(M_{gc}^{\text{com}})$, Eq. 8, and a non-smooth component $R(M_{gc}^{\text{com}})$, Eq. 9. The smooth part is

$$F\big(M_{gc}^{\text{com}}\big) \;=\; \big\| P_{\Omega_{gc}}\big(M_{gc} - M_{gc}^{\text{com}}\big)\big\|_F^2, \quad (22)$$

whose gradient satisfies

$$\nabla F(M_{gc}^{\text{com}}) \;=\; -2\, P_{\Omega_{gc}}\Big(M_{gc} - M_{gc}^{\text{com}}\Big). \quad (23)$$

Because $P_{\Omega_{gc}}$ is a linear (masking) operator, $\nabla F(\cdot)$ is Lipschitz continuous. Formally, there exists $L > 0$ such that

$$\|\nabla F(X) - \nabla F(Y)\|_F \;\le\; L\,\|X - Y\|_F, \forall X, Y. \quad (24)$$

This L-smoothness property is fundamental for analyzing the convergence of proximal gradient-type methods.

The non-smooth part is

$$R\big(M_{gc}^{\text{com}}\big) \;=\; \lambda_1 \big\| P_{\chi^\tau}\big(M_{gc}^{\text{com}}\big)\big\|_*, \quad (25)$$

where $\|\cdot\|_*$ denotes the nuclear norm. The nuclear norm is convex, with its proximal operator given by Singular Value Thresholding (SVT). Since P_{χ^τ} is an elementwise mask, the operator $M_{gc}^{\text{com}} \mapsto P_{\chi^\tau}(M_{gc}^{\text{com}})$ remains linear and contractive, and thus the composition $\big\|P_{\chi^\tau}(\cdot)\big\|_*$ is likewise convex and admits a closed-form proximal operator. This ensures that the non-smooth term $R(\cdot)$ is efficiently handled within a forward-backward splitting scheme.

In addition to the proximal step, we impose the linear constraints $f\big(A_{ij}, C_{ij}(M_{gc}^{\text{com}})\big) \;\ge\; \alpha$, $\forall (i,j) \in \text{TopX}(i)$, which define the set

$$\mathcal{S} = \Big\{ M_{gc}^{\text{com}} \in \mathbb{R}^{m\times n} \;\Big|\; f\big(A_{ij}, C_{ij}\big(M_{gc}^{\text{com}}\big)\big) \ge \alpha \quad \forall (i,j) \in \text{TopX}(i) \Big\}, \quad (26)$$

Because these constraints are linear in the entries of M_{gc}^{com}, the set $\mathcal{S}$ is a closed, convex polyhedron. After each proximal update, we project the intermediate estimate onto $\mathcal{S}$ by solving a convex quadratic program, which maintains feasibility of the iterates.

From classical results in convex analysis (Beck and Teboulle, 2009; Combettes and Pesquet, 2011; Bauschke and Combettes, 2017), it follows that if $F(\cdot)$ is convex with an L-Lipschitz continuous gradient and $R(\cdot)$ is convex, the forward-backward splitting method converges to a global minimizer of $F + R$. When combined with a projection step onto a closed, convex set $\mathcal{S}$, one can view the projection as the proximal operator of the indicator function $\delta_{\mathcal{S}}(\cdot)$, which preserves the global convergence guarantees. Consequently, under mild assumptions (e.g., finite entries and bounded parameters), the sequence $\{M_{gc}^{\text{com},(t)}\}$ converges to a global optimum of Eq. 1. Thus, the proposed method is not only computationally tractable but also theoretically sound, which ensures convergence to a robust and interpretable completed gene-cancer matrix.

Cancer Data	Time (in sec)	Time (GPU hours)
Prostate Cancer	358.70	0.099638889
Cervical Cancer	183.30	0.050916667
Breast Cancer	360.07	0.100019444
Lung Cancer	1521.37	0.4226027778
Sarcoma Cancer	14.10	0.003916667

Table 5: Training time of **LILY** on datasets collected for each cancer type.

D Training Time Analysis

We trained our model on a single Tesla V-100 GPU with 16GB of CUDA memory. Table 5 details the training time for data collected by each cancer type. Specifically, the prostate cancer dataset comprises data from 10,000 oncogenomics articles, 25,524 relevant lines, 2,965 annotated genes, and 990 cancer types; the cervical cancer dataset includes 10,000 articles, 17,115 relevant lines, 2,407 annotated genes, and 692 cancer types; the breast cancer dataset is based on 10,000 articles, 38,620 relevant lines, 3,641 annotated genes, and 829 cancer types; the lung cancer dataset consists of 10,000 articles, 60,532 relevant lines, 6,242 annotated genes, and 1,716 cancer types; and the sarcoma cancer dataset is derived from 1,061 articles, 5,679 relevant lines, 679 annotated genes, and 283 cancer types.