Detecting High-Potential SMEs with Heterogeneous Graph Neural Networks

Yijiashun Qi*
elijahqi@umich.edu
University of Michigan
USA

Hanzhe Guo
hanzheg@umich.edu
University of Michigan
USA

Yijiazhen Qi
qiyijiazhen@gmail.com
The University of Hong Kong
Hong Kong, China

Abstract

Small and Medium Enterprises (SMEs) constitute the vast majority of businesses globally and generate a significant portion of economic activity, yet systematically identifying high-potential SMEs remains an open challenge. We introduce SME-HGT, a Heterogeneous Graph Transformer framework that predicts which early-stage (Phase I) innovation grant awardees will advance to Phase II funding using exclusively public data. We construct a heterogeneous graph with 32,268 company nodes, 124 research topic nodes, and 13 funding agency nodes connected by approximately 99,000 edges across five semantic relation types. Against baselines including MLP and R-GCN, XGBoost achieves the strongest overall performance (AUPRC 0.629 ± 0.003, AUROC 0.652 ± 0.003), consistent with the documented strength of tree-based methods on tabular data. However, graph-based methods substantially outperform the non-graph MLP baseline (SME-HGT AUROC 0.646 ± 0.003 vs. MLP 0.634 ± 0.014), confirming that relational structure provides meaningful signal for neural prediction. Feature ablation identifies Phase I award count as the dominant predictive feature (ΔAUPRC = −0.026 when removed); temporal robustness analysis across three non-overlapping time windows reveals substantial concept drift (10–18 pp AUPRC decline), motivating periodic retraining. At a screening depth of 100 companies, XGBoost attains 92.4% precision (2.21× lift), while SME-HGT achieves 83.4% (2.00× lift). Our temporal evaluation protocol prevents information leakage, and our reliance on public data ensures reproducibility. These results demonstrate that relational structure among firms, research topics, and funding bodies provides meaningful signal for SME potential assessment, with implications for economic policymakers and early-stage investors globally.

CCS Concepts

• **Computing methodologies** → **Neural networks**; *Learning latent representations*; • **Information systems** → *Graph-based database models*; • **Applied computing** → *Economics*.

*Corresponding author.

Conference'17, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Keywords

graph neural networks, heterogeneous graphs, SME identification, innovation grants, economic analytics

ACM Reference Format:
Yijiashun Qi, Hanzhe Guo, and Yijiazhen Qi. 2026. Detecting High-Potential SMEs with Heterogeneous Graph Neural Networks. In . ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Small and Medium Enterprises (SMEs) are the primary engine of innovation and economic growth globally. According to international economic development reports [24], SMEs account for the overwhelming majority of all businesses, employ a large segment of the private workforce, and generate a substantial portion of global economic activity. Identifying which of these firms have the highest growth potential is a longstanding challenge for economic analysts, investors, and grant program administrators alike.

National and regional innovation grant programs are major sources of early-stage technology funding, distributing billions annually [15]. These programs typically operate in sequential phases: Phase I awards fund feasibility studies, while Phase II awards support full R&D. Phase I to Phase II progression is rigorously evaluated and serves as a proxy for technical and commercial potential [1, 16]. Predicting which Phase I awardees advance to Phase II is both a practical policy question and a test case for data-driven assessment.

Current approaches to SME assessment rely predominantly on tabular features—financial metrics, patent counts, employee size—or manual expert review. These methods miss an important dimension: the *relational structure* in which firms are embedded. A company's potential may be signaled not only by its own attributes but also by its connections to specific research domains, its funding relationships with sponsoring agencies, and its proximity to other successful firms in the innovation ecosystem.

Graph Neural Networks (GNNs) offer a natural framework for capturing such relational information. Recent work has demonstrated the power of GNNs for tasks ranging from social network analysis [12] to molecular property prediction [10]. Broader artificial intelligence methodologies are similarly transforming critical physical infrastructure, optimizing design and control strategies in power electronics and energy systems [7]. In the economic domain, heterogeneous graph methods have shown promise for firm-level prediction tasks [13], though prior work has typically relied on proprietary data sources.

In this paper, we make four contributions:

(1) **Public-data heterogeneous graph.** We construct a multi-relational heterogeneous graph of an innovation grant ecosystem using exclusively public database records, with three

node types (companies, research topics, funding agencies) and five edge types capturing operational, funding, co-topic, co-agency, and co-state relationships.

(2) **SME-HGT framework.** We adapt the Heterogeneous Graph Transformer [13] for SME potential prediction, demonstrating that graph-based methods significantly outperform non-graph neural baselines and providing an honest comparison with XGBoost [3], which achieves the highest overall performance.

(3) **Comprehensive empirical analysis.** We provide feature and edge-type ablation studies, multi-edge vs. single-edge comparison, temporal robustness evaluation across three time windows, and detailed error analysis with case studies.

(4) **Temporal evaluation protocol.** We design a strict temporal split that prevents information leakage, using feature cutoffs, label horizons, and chronological train/validation/test partitions to simulate realistic deployment conditions.

2 Related Work

2.1 Graph Neural Networks for Economic Applications

Graph-based learning has been increasingly applied to economic and financial prediction tasks. Kipf and Welling [14] introduced Graph Convolutional Networks (GCNs) for semi-supervised node classification, and subsequent architectures including GAT [25] and GraphSAGE [12] have been applied to fraud detection, credit scoring, and supply chain analysis. Zhou et al. [27] provide a comprehensive survey of GNN architectures and applications. Recent work has also explored GNN-driven hierarchical mining strategies for complex imbalanced datasets [20], a challenge relevant to our setting where Phase II recipients constitute a minority class.

In the context of firm-level prediction, GNNs can leverage corporate networks, ownership structures, and industry co-membership graphs to augment tabular features. However, most economic applications of GNNs operate on homogeneous graphs and rely on proprietary data. Our work addresses both limitations: we use a heterogeneous graph that explicitly models different entity types and their distinct relations, and we rely exclusively on publicly available grant program data.

2.2 Innovation Grant Analytics

Innovation grant programs have been extensively studied in the economics literature. Lerner [15] conducted a foundational analysis showing that grant awardees exhibit significantly higher growth rates than matched non-awardees. Link and Scott [16] examined the role of these programs in bridging the "valley of death" between basic research and commercialization. Audretsch et al. [1] studied the relationship between early-stage funding and small firm innovation output. Toole and Czarnitzki [23] analyzed phase progression specifically in the biomedical sector.

Despite this rich literature, existing grant analyses are predominantly retrospective and descriptive. To our knowledge, no prior work has applied graph-based machine learning to predict Phase II progression using the relational structure of the innovation ecosystem.

2.3 Heterogeneous Graph Neural Networks

Real-world graphs frequently contain multiple node and edge types. Several architectures have been proposed for heterogeneous graph learning. Schlichtkrull et al. [21] introduced R-GCN with relation-specific weight matrices. Wang et al. [26] proposed Heterogeneous Graph Attention Network (HAN) using metapath-based aggregation, building on the metapath concept from Sun et al. [22] and Dong et al. [8]. Most relevant to our work, Hu et al. [13] developed the Heterogeneous Graph Transformer (HGT), which uses type-specific linear projections and multi-head attention without requiring predefined metapaths.

We adopt HGT as our primary architecture due to its flexibility in handling arbitrary heterogeneous schemas and its ability to learn type-dependent attention patterns, which we hypothesize are important for capturing the distinct roles of companies, research topics, and funding agencies.

3 Methodology

3.1 Data

We use publicly available data from an open innovation awards database, which contains records of public grants made by participating funding bodies. The raw dataset comprises 207,726 award records spanning multiple decades. Each record includes the company name, award amount, award year, program phase (I or II), awarding agency, and research topic code.

We apply the following data cleaning pipeline: (1) standardize company names by converting to uppercase, removing legal suffixes, and collapsing whitespace for entity resolution; (2) parse phase numbers from heterogeneous string formats; (3) extract research topic prefixes and agency identifiers; and (4) remove records with missing critical fields. After entity resolution, the cleaned dataset contains 32,268 unique companies.

3.2 Graph Construction

We construct a heterogeneous graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with three node types and five semantic edge types. Reverse edges are automatically added for bipartite relations (operates_in, awarded_by), while co-occurrence edges (co_topic, co_agency, co_state) are stored as undirected, yielding approximately 99,000 total edges. Table 1 summarizes the graph schema.

Company tabular features (7 dimensions) characterize each firm's Phase I history:

(1) **Log total Phase I funding** — cumulative funding reflects investor confidence [15].
(2) **Log Phase I award count** — repeat awardees demonstrate sustained competitiveness [1].
(3) **Agency diversity** — number of unique agencies; breadth signals versatility [16].
(4) **Years active** — program duration captures persistence.
(5) **Log avg. Phase I award size** — larger awards may indicate more ambitious projects.
(6) **Topic diversity** — number of unique research topics; breadth may signal commercialization potential.
(7) **Award recency** — normalized most recent Phase I year; recent activity signals engagement.

Table 1: Heterogeneous Graph Schema. *7 tabular + 383 text embedding dims.

Node Type	Count	Features
Company	32,268	390*
Topic	124	2
FundingAgency	13	3
Edge Types		
(Company, OPERATES_IN, Topic)		
(Company, AWARDED_BY, FundingAgency)		
(Company, CO_TOPIC, Company)		
(Company, CO_AGENCY, Company)		
(Company, CO_STATE, Company)		

All features are min-max normalized to $[0, 1]$. Additionally, we concatenate 383-dimensional sentence-transformer embeddings of company names, yielding a total input dimension of 390 per company. These embeddings are L2-normalized and capture lexical similarity between firm identities. All models—including XGBoost and MLP—receive the full 390-dimensional feature vector. We validate each tabular feature's contribution through an ablation study (Section 4.5).

Topic features (2 dimensions) characterize each research area: log number of participating companies and log total awards within the topic, both normalized.

Funding agency features (3 dimensions): log award count, log total funding amount, and log average award size, all normalized.

Edge construction. For bipartite relations, each company is connected to *all* research topics and funding agencies associated with its Phase I awards, not only the primary (modal) one. This multi-edge construction captures the full breadth of a firm's engagement: a company with Phase I awards across three agencies receives three AWARDED_BY edges, and similarly for OPERATES_IN edges to research topics. For company–company co-occurrence edges, firms sharing the same primary research topic (CO_TOPIC), the same primary agency (CO_AGENCY), or the same state (CO_STATE) are connected, with a cap of 50 edges per group and 20 per node to manage density. We compare this multi-edge construction against a single-edge baseline (primary only) in Section 4.6.

Temporal integrity. All company features are computed using only Phase I awards prior to the feature cutoff date (January 1, 2018). Phase II information is never included in node features, ensuring strict separation between inputs and prediction targets.

3.3 Label Design

We define the prediction target as a binary variable: whether a Phase I awardee receives a Phase II award within a 5-year horizon of their first Phase I award. Formally, for company c:

$$y_c = \mathbb{1}\left[\exists \text{ Phase II award for } c \text{ with year} \le t_c^{(1)} + 5\right] \tag{1}$$

where $t_c^{(1)}$ denotes the year of company c's first Phase I award.

We partition companies into temporally non-overlapping sets based on their first Phase I award year:

- **Train:** first Phase I year < 2018
- **Validation:** $2018 \le$ first Phase I year < 2020 $(n = 2{,}353)$
- **Test:** $2020 \le$ first Phase I year < 2022 $(n = 2{,}689)$

Companies with first Phase I year ≥ 2022 are excluded due to insufficient observation window for the 5-year label horizon. The test set positive rate is 41.8% ($n_+ = 1{,}124$), reflecting moderate class balance.

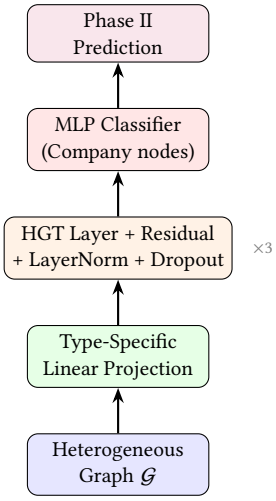

Figure 1: SME-HGT architecture. Input features are projected per node type to a shared dimension, then refined through three HGT layers with residual connections. Only company node embeddings are passed to the classifier.

3.4 SME-HGT Architecture

Our model, SME-HGT, consists of three components: type-specific input projection, heterogeneous graph transformer layers, and a classification head. Figure 1 illustrates the overall architecture.

Input projection. Each node type τ receives a dedicated linear transformation $\mathbf{h}_v^{(0)} = \mathbf{W}_\tau \mathbf{x}_v + \mathbf{b}_\tau$, mapping heterogeneous input features to a shared hidden dimension $d = 128$.

HGT layers. We stack $L = 3$ HGT [13] layers. Each layer computes type-dependent attention weights across all edge types using $H = 4$ heads. For target node t receiving messages from source s via relation r:

$$\alpha_{s,r,t} = \underset{s \in \mathcal{N}_r(t)}{\text{Softmax}} \frac{(\mathbf{W}^Q_{\tau(t)} \mathbf{h}_t)^\top (\mathbf{W}^K_{\tau(s),r} \mathbf{h}_s)}{\sqrt{d/H}} \tag{2}$$

$$\mathbf{h}_t^{(\ell+1)} = \sum_{r \in \mathcal{R}} \sum_{s \in \mathcal{N}_r(t)} \alpha_{s,r,t} \cdot \mathbf{W}^V_{\tau(s),r} \mathbf{h}_s^{(\ell)} \tag{3}$$

Each HGT layer is followed by a residual connection, per-type layer normalization [2], and dropout ($p = 0.2$).

Classification head. After L layers of message passing, company node representations are passed through a two-layer MLP classifier:

$$\hat{y}_c = \text{Linear}_{64 \to 2}\Big(\text{Dropout}\Big(\text{ReLU}\Big(\text{Linear}_{128 \to 64}(\mathbf{h}_c^{(L)})\Big)\Big)\Big) \tag{4}$$

Table 2: Test Set Results (mean ± std over 10 seeds)

Model	AUPRC	AUROC	F1
XGBoost	**0.629**$_{\pm.003}$	**0.652**$_{\pm.003}$	0.590$_{\pm.000}$
MLP	0.591$_{\pm.025}$	0.634$_{\pm.014}$	**0.593**$_{\pm.001}$
R-GCN	0.598$_{\pm.006}$	0.648$_{\pm.006}$	0.593$_{\pm.002}$
SME-HGT	0.606$_{\pm.005}$	0.646$_{\pm.003}$	0.591$_{\pm.001}$

producing logits for binary classification. Topic and agency node embeddings are not directly classified; they serve as context through message passing.

3.5 Training Procedure

We optimize all models using AdamW [18] with learning rate 10^{-3} and weight decay 5×10^{-4}. The learning rate follows a linear warmup over 10 epochs followed by cosine annealing [17] over the remaining epochs. Training proceeds for a maximum of 200 epochs with early stopping based on validation loss (patience = 30). Gradient norms are clipped at 1.0. The loss function is binary cross-entropy, appropriate given the near-balanced class distribution in our dataset. All experiments are implemented using PyTorch Geometric [9].

4 Experiments

4.1 Experimental Setup

Baselines. We compare SME-HGT against three baselines:

- **XGBoost** [3]: A gradient-boosted tree ensemble operating on company node features only. This strong tabular baseline uses 500 estimators, max depth 6, and early stopping on validation loss. XGBoost represents the state of the art for tabular prediction and has been shown to outperform deep learning on medium-sized structured datasets [11].
- **MLP**: A 3-layer feedforward network operating on company node features only, ignoring all graph structure. Each layer applies Linear → ReLU → BatchNorm → Dropout with hidden dimension 128. This baseline isolates the contribution of graph structure by using identical features.
- **R-GCN** [21]: A relational graph convolutional network using SAGEConv [12] layers converted to heterogeneous form via per-type input projections. It uses 2 message-passing layers with residual connections and LayerNorm, followed by the same classification head.

All neural models share the same hidden dimension (128), dropout rate (0.2), and training hyperparameters. Model sizes: SME-HGT 729,863 parameters; R-GCN 340,674; MLP 18,178; XGBoost is non-parametric. Results are averaged over 10 seeds.

Metrics. We report AUPRC (Area Under the Precision-Recall Curve) as the primary metric [6]. We additionally report AUROC, F1 at the optimal threshold, Precision@K for $K \in \{100, 500, 1000\}$, and Lift@K.

4.2 Main Results

Table 2 presents the primary experimental results on the held-out test set.

Table 3: Ranking Metrics on Test Set (mean over 10 seeds)

	Precision@K			Lift@K		
Model	@100	@500	@1K	@100	@500	@1K
XGBoost	**.924**	**.711**	**.578**	**2.21**	**1.70**	**1.38**
MLP	.845	.668	.550	2.02	1.60	1.32
R-GCN	.826	.673	.572	1.98	1.61	1.37
HGT	.834	.689	.576	2.00	1.65	1.38

XGBoost achieves the highest performance across all three metrics: AUPRC 0.629 ± 0.003, AUROC 0.652 ± 0.003, consistent with the documented strength of gradient-boosted trees on medium-sized tabular data [11]. Among neural methods, graph-based architectures substantially outperform the non-graph MLP: SME-HGT improves over MLP by +1.5 pp on AUPRC and +1.2 pp on AUROC, while R-GCN improves by +0.7 pp and +1.4 pp respectively. HGT and R-GCN achieve comparable AUROC (0.646 vs. 0.648), but HGT exhibits lower variance (±0.003 vs. ±0.006). F1 scores are comparable (∼0.590–0.593) across all models.

The XGBoost advantage reflects the strength of tree-based methods for probability estimation on tabular features, particularly with rich text embeddings. However, graph-based methods significantly improve over MLP—which uses identical features without graph structure—confirming that relational context provides signal for neural learning. The GNN approach offers complementary advantages: (1) relational extensibility as new node types become available; (2) interpretable attention patterns over semantic edge types; and (3) a natural framework for heterogeneous data.

4.3 Ranking Performance

Table 3 presents Precision@K and Lift@K metrics, which are most relevant for practical screening applications where analysts review a fixed number of top-ranked candidates.

XGBoost achieves the highest ranking performance at all screening depths, with 92.4% Precision@100 and 71.1% Precision@500. Among neural methods, SME-HGT achieves the best mid-range ranking: 68.9% Precision@500 compared to 67.3% for R-GCN and 66.8% for MLP, confirming that graph structure aids candidate prioritization beyond tabular features alone. At Precision@1000, HGT (57.6%) nearly matches XGBoost (57.8%), with R-GCN (57.2%) close behind. All models substantially outperform random selection (Lift > 1.0), with XGBoost achieving the highest lift at shallow depths (2.21× at top 100).

4.4 Analysis

Graph structure aids neural prediction. While XGBoost achieves the highest performance on all metrics, graph-based methods (HGT, R-GCN) consistently outperform the non-graph MLP baseline on AUPRC (+1.5 pp), AUROC (+1.2 pp), and Precision@500 (+2.1 pp). This indicates that relational structure in the innovation ecosystem provides meaningful signal for neural models, even though tree-based methods capture similar patterns through feature interactions alone. Edge ablation (Section 4.5) confirms that the benefit

Table 4: Feature Ablation (HGT, AUPRC, mean ± std over 5 seeds). Larger Δ = more important feature.

Removed Feature	AUPRC	Δ vs. Full
None (full model)	$0.603_{\pm.002}$	—
− log total Phase I	$0.597_{\pm.002}$	−0.006
− log Phase I count	$0.577_{\pm.003}$	−0.026
− agency diversity	$0.604_{\pm.002}$	+0.001
− years active	$0.607_{\pm.002}$	+0.005
− log avg Phase I size	$0.601_{\pm.002}$	−0.002
− topic diversity	$0.603_{\pm.002}$	+0.000
− award recency	$0.603_{\pm.003}$	+0.000

arises primarily from bipartite edges linking companies to topics and agencies.

HGT offers superior stability. While R-GCN achieves comparable mean AUROC, its AUPRC variance (±0.006) exceeds that of SME-HGT (±0.005), and its AUROC variance (±0.006) is double HGT's (±0.003). MLP shows the highest neural variance (±0.025 AUPRC, ±0.014 AUROC), suggesting that graph-based aggregation stabilizes predictions. XGBoost exhibits the lowest variance (±0.003), as expected for ensemble methods.

Convergence behavior. SME-HGT converges in 23–41 epochs across seeds (mean training time ∼149 s), while R-GCN requires 26–77 epochs (∼31 s) and MLP 50–56 epochs (∼8 s). Despite the higher per-epoch cost of attention-based message passing, all training times remain within practical bounds for an offline screening application.

Moderate overall discrimination. We note that absolute AUROC values (0.63–0.65) indicate moderate rather than strong discriminative power. This reflects the inherent difficulty of predicting Phase II outcomes—which depend on proposal quality, reviewer composition, and evolving funding priorities—from structural and historical features alone. Nevertheless, the ranking metrics demonstrate substantial practical value: XGBoost screening the top 100 candidates yields 92.4% precision, a 2.21× lift over random selection.

4.5 Feature and Edge Ablation

To quantify the marginal contribution of each input feature and edge type, we conduct systematic ablation studies. For *feature ablation*, we zero out one of the 7 company features at a time and retrain SME-HGT (×5 seeds). For *edge-type ablation*, we remove one of the 5 edge types from the graph and retrain.

Feature ablation. Table 4 reports test-set AUPRC when each tabular feature is zeroed out. Log Phase I count is the dominant feature ($\Delta = -0.026$), indicating that repeat-award history is the strongest tabular signal. Log total Phase I amount is second ($\Delta = -0.006$). The remaining five features show marginal or negligible effects ($|\Delta| \leq 0.005$), with years active and agency diversity slightly improving when removed—suggesting possible redundancy with other features or mild noise. Note that these ablations affect only the 7 tabular dimensions; the 383 text embedding dimensions remain intact in all conditions.

Edge-type ablation. Table 5 reports test-set AUPRC when each edge type is excluded. The bipartite edges—AWARDED_BY ($\Delta = -0.009$) and OPERATES_IN ($\Delta = -0.007$)—are the most beneficial: removing either degrades performance, confirming that direct company-to-agency and company-to-topic links provide meaningful relational signal. Surprisingly, removing CO_TOPIC edges *improves* AUPRC by +0.012, suggesting that broad topic-based company-company connections introduce noise that dilutes informative message passing. The CO_AGENCY and CO_STATE edges have negligible effects ($|\Delta| \leq 0.002$). These results suggest that the graph's value comes primarily from heterogeneous bipartite structure (companies linked to topics and agencies) rather than homogeneous company-company co-occurrence edges.

Table 5: Edge-Type Ablation (HGT, AUPRC, mean ± std over 5 seeds).

Removed Edge Type	AUPRC	Δ vs. Full
None (full model)	$0.603_{\pm.002}$	—
− OPERATES_IN	$0.596_{\pm.003}$	−0.007
− AWARDED_BY	$0.594_{\pm.005}$	−0.009
− CO_TOPIC	$0.615_{\pm.002}$	+0.012
− CO_AGENCY	$0.605_{\pm.005}$	+0.002
− CO_STATE	$0.601_{\pm.006}$	−0.002

Table 6: Multi-Edge vs. Single-Edge (AUPRC, mean ± std over 5 seeds).

Model	Edge Mode	AUPRC	AUROC
SME-HGT	Multi	$0.606_{\pm.003}$	$0.647_{\pm.003}$
	Single	$0.599_{\pm.004}$	$0.645_{\pm.004}$
R-GCN	Multi	$0.607_{\pm.007}$	$0.653_{\pm.005}$
	Single	$0.594_{\pm.010}$	$0.650_{\pm.007}$
MLP	Multi	$0.557_{\pm.070}$	$0.608_{\pm.051}$
	Single	$0.557_{\pm.070}$	$0.608_{\pm.051}$
XGBoost	Multi	$\mathbf{0.628}_{\pm.005}$	$\mathbf{0.653}_{\pm.004}$
	Single	$0.628_{\pm.005}$	$0.653_{\pm.004}$

4.6 Multi-Edge vs. Single-Edge Comparison

Our graph construction links each company to *all* associated topics and agencies (multi-edge), rather than only the primary one (single-edge). Table 6 compares both modes across all four models.

Multi-edge construction yields a consistent improvement for graph-based models (HGT: +0.7 pp AUPRC; R-GCN: +1.3 pp AUPRC), while MLP and XGBoost are unaffected since they do not use edges. This confirms that the richer multi-edge connectivity provides additional relational signal beyond the primary association. The benefit is larger for R-GCN, suggesting that its relation-specific convolutions can better exploit the additional edge diversity.

4.7 Temporal Robustness

To assess robustness to concept drift, we evaluate all models across three non-overlapping temporal windows (Table 7).

Table 7: Temporal Robustness (AUPRC, mean ± std over 5 seeds). Test periods: Early (2018–2020), Current (2020–2022), Late (2021–2023).

Model	Early	Current	Late
XGBoost	**0.732**±.001	**0.628**±.005	**0.548**±.005
MLP	0.660±.078	0.557±.070	0.433±.086
R-GCN	0.676±.059	0.607±.007	0.483±.080
SME-HGT	0.717±.004	0.606±.003	0.542±.005

All models show substantial AUPRC decline from Early to Late windows. XGBoost degrades from 0.732 to 0.548 (−18.4 pp), while SME-HGT drops from 0.717 to 0.542 (−17.5 pp). MLP and R-GCN show even larger absolute drops with high variance (MLP: 0.660 → 0.433; R-GCN: 0.676 → 0.483), suggesting that graph structure provides a stabilizing effect under distribution shift.

Two factors drive the observed decline: (1) genuine concept drift as the innovation ecosystem evolves (changing agency priorities, emerging research topics, shifting funding patterns); and (2) right-censoring in the Late window, where the 5-year label horizon extends beyond available data, potentially mislabeling companies that will eventually receive Phase II. SME-HGT and XGBoost maintain notably low variance across seeds (±0.003–0.005), while MLP and R-GCN exhibit high instability in non-standard windows (±0.059–0.086). These results motivate annual retraining and careful monitoring of validation performance in deployment.

4.8 Error Analysis

We analyze prediction errors from the best SME-HGT checkpoint on the test set to understand failure modes and identify areas for improvement.

Confusion matrix. At threshold $\tau = 0.5$, the test set ($n = 2{,}689$) decomposes into: TP = 620, FP = 494, FN = 504, TN = 1,071. The false positive rate (FP/(FP+TN) = 31.6%) is lower than the false negative rate (FN/(FN+TP) = 44.8%), indicating a tendency toward conservative prediction at this threshold. The F1-optimal threshold ($\tau \approx 0.19$) yields extremely high recall (99.8%) at low precision (42%), reflecting that probability scores are concentrated near the decision boundary—reinforcing that the model is best used as a ranker rather than a binary classifier.

Errors by agency. At $\tau = 0.5$, error rates vary substantially by funding agency. DOD (the largest agency, $n = 1{,}175$) shows the highest FPR (45%) but lowest FNR (32%), suggesting the model overpredicts for defense-sector firms. NSF ($n = 463$) exhibits the opposite: very low FPR (11%) but high FNR (92%), meaning the model rarely predicts NSF-funded firms as positive. HHS and DOE show intermediate patterns (FPR 30–33%, FNR 47–57%). These agency-specific biases likely reflect differences in Phase II transition rates across programs.

Errors by company profile. False negatives disproportionately involve single-award companies: 437 of 504 FN (87%) have only one Phase I award, compared to 67% single-award in the test set. These firms have minimal relational context for the GNN. False positives tend to be well-connected multi-award firms (3–5 Phase I awards) that the model identifies as structurally similar to successful companies but that did not receive Phase II within the observation window.

Case studies. Among the top false positives (highest predicted probability, $y = 0$), we observe multi-award DOD-funded companies with 2–5 Phase I awards across 3–4 years that did not receive Phase II. These are plausible candidates whose applications may have been rejected for factors outside our features (proposal quality, reviewer composition), or who pursued alternative paths (direct contracts, STTR). Among the top false negatives (lowest predicted probability, $y = 1$), we find exclusively single-award, single-year companies in niche agencies—the sparse-context scenario where GNN message passing provides minimal benefit.

5 Discussion

5.1 Policy Implications

Our results have direct relevance for global grant program administration and small business policy. A screening tool achieving 92.4% precision among its top 100 predictions (XGBoost) could substantially reduce the manual review burden for program managers. At a base rate of 41.8%, random selection would yield approximately 42 successful firms out of 100 reviewed; the best model identifies approximately 92. This 2.21× lift translates to more efficient allocation of expert review resources and potentially earlier identification of high-potential firms for targeted technical assistance.

The public-data-only design is a deliberate choice: any jurisdiction with structured grant or subsidy data can construct similar heterogeneous graphs to assess participant potential without requiring proprietary databases. This makes our approach replicable for analogous programs across various national innovation systems.

5.2 Concept Drift and Retraining

Our temporal robustness analysis (Section 4.7) reveals a modest but consistent decline in performance across time windows, consistent with distributional shift in the innovation ecosystem. Funding priorities evolve, new research topics emerge, and agency review criteria change over time. This concept drift is not unique to our setting—it is a well-documented challenge in deployed ML systems [19].

The observed degradation is substantial (10–18 pp AUPRC from Early to Late), though partially attributable to right-censoring in the Late window. We recommend annual graph reconstruction and model retraining to incorporate the latest Phase I data, with monitoring of validation AUPRC to trigger additional updates if performance drops below acceptable thresholds. The fully automated data pipeline (download, clean, build graph, train) makes such retraining operationally feasible. Notably, SME-HGT and XGBoost maintain low cross-seed variance even in shifted windows, while MLP and R-GCN become highly unstable, suggesting that both attention-based graph aggregation and ensemble methods are more robust to distribution shift.

5.3 Outcome Measure Validity

We use Phase II receipt as a proxy for SME potential, following prior work that establishes Phase I→II progression as a meaningful indicator of technical and commercial viability [1, 16]. Phase II awards involve rigorous technical review and require demonstration of

Phase I feasibility, making them a higher bar than Phase I selection alone. Toole and Czarnitzki [23] further validate this proxy by showing Phase II receipt correlates with subsequent patent activity and firm growth.

However, Phase II receipt is an imperfect measure of long-term potential. Some high-potential firms may pursue alternative funding paths (venture capital, direct federal contracts, STTR) rather than Phase II. Conversely, some Phase II recipients may fail to commercialize their research. A richer outcome measure might combine Phase II receipt with downstream indicators such as patent filings, revenue growth, or acquisition events. We leave this multi-objective formulation to future work, noting that it requires linking additional proprietary datasets (e.g., USPTO patents, Crunchbase) that would compromise our public-data-only design.

5.4 Limitations

Several limitations warrant discussion.

(1) Our graph relies exclusively on historical grant program data; incorporating additional public sources such as patent filings or local economic data could strengthen both node features and graph connectivity.
(2) We observe only Phase I survivors—companies that already won competitive grants—introducing selection bias relative to the broader SME population.
(3) The moderate overall discrimination (AUROC ≈ 0.65) suggests that important predictive signals—proposal text quality, team composition, technology readiness level—are not captured in our current structural features.
(4) Our entity resolution relies on rule-based string matching, which may fail on name variants or subsidiaries not covered by our normalization rules.
(5) While XGBoost outperforms GNNs on AUPRC, it cannot leverage relational structure; future work should investigate hybrid approaches combining tree-based feature processing with graph-based relational reasoning.

5.5 Future Work

Several directions merit exploration: (1) incorporating textual features from award abstracts via language model embeddings and retrieval-augmented analysis methods [4, 5] to capture proposal quality signals; (2) adding temporal dynamics through time-aware graph architectures such as evolving GCNs; (3) developing hybrid models that combine XGBoost's tabular strength with GNN relational features; (4) extending to downstream prediction targets such as patent filing, revenue growth, or acquisition events; and (5) developing an interactive dashboard for program managers to explore model predictions alongside company profiles.

6 Conclusion

We presented SME-HGT, a Heterogeneous Graph Transformer for identifying high-potential small businesses using public grant data. By constructing a heterogeneous graph with 32,268 companies, 124 topics, and 13 agencies across five edge types, we showed that relational structure provides meaningful signal beyond tabular features.

XGBoost achieves the highest performance across all metrics (AUPRC 0.629, AUROC 0.652), while graph-based methods (HGT, R-GCN) significantly outperform the non-graph MLP baseline, confirming that relational structure provides meaningful signal for neural prediction. Feature ablation reveals Phase I award count as the dominant predictive signal, while edge ablation shows that bipartite company-topic and company-agency edges drive the GNN's advantage. Temporal analysis across three windows shows substantial concept drift (10–18 pp AUPRC), motivating annual retraining.

At a practical screening depth of 100 candidates, XGBoost attains 92.4% precision (2.21× lift), while SME-HGT achieves 83.4% (2.00× lift). Our strict temporal evaluation protocol and exclusive reliance on public data ensure that these results reflect realistic deployment conditions. This work illustrates the potential of heterogeneous graph neural networks for policy-relevant prediction tasks in the global SME innovation ecosystem.

Acknowledgments

Data sourced from public open-data grant portals. Computational experiments were conducted using PyTorch and PyTorch Geometric.

References

[1] David B. Audretsch, Albert N. Link, and John T. Scott. 2002. Public/Private Technology Partnerships: Evaluating SBIR-Supported Research. *Research Policy* 31, 1 (2002), 145–158.
[2] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. *arXiv preprint arXiv:1607.06450* (2016).
[3] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 785–794.
[4] Zhiyuan Cheng, Longying Lai, and Yue Liu. 2026. Resolving the Robustness-Precision Trade-off in Financial RAG through Hybrid Document-Routed Retrieval. arXiv:2603.26815 [cs.CL] https://arxiv.org/abs/2603.26815
[5] Zhiyuan Cheng, Longying Lai, Yue Liu, Kai Cheng, and Xiaoxi Qi. 2026. Enhancing Financial Report Question-Answering: A Retrieval-Augmented Generation System with Reranking Analysis. arXiv:2603.16877 [cs.CL] https://arxiv.org/abs/2603.16877
[6] Jesse Davis and Mark Goadrich. 2006. The Relationship Between Precision-Recall and ROC Curves. In *International Conference on Machine Learning (ICML)*. 233–240.
[7] T. K. Ding, D. Xiang, T. Sun, Y. Qi, Z. Zhao, and Y. Qi. 2025. Artificial intelligence applications in power electronics. In *2025 IEEE 7th International Conference on Energy Systems and Electrical Power*. IEEE.
[8] Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable Representation Learning for Heterogeneous Networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 135–144.
[9] Matthias Fey and Jan Eric Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric. In *ICLR Workshop on Representation Learning on Graphs and Manifolds*.
[10] Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. 2017. Neural Message Passing for Quantum Chemistry. In *International Conference on Machine Learning (ICML)*. 1263–1272.
[11] Léo Grinsztajn, Edouard Oyallon, and Gaël Varoquaux. 2022. Why do Tree-Based Models Still Outperform Deep Learning on Typical Tabular Data?. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35.
[12] William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive Representation Learning on Large Graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 30.
[13] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous Graph Transformer. In *Proceedings of The Web Conference (WWW)*. 2704–2710.
[14] Thomas N. Kipf and Max Welling. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *International Conference on Learning Representations (ICLR)*.
[15] Josh Lerner. 1999. The Government as Venture Capitalist: The Long-Run Impact of the SBIR Program. *The Journal of Business* 72, 3 (1999), 285–318.
[16] Albert N. Link and John T. Scott. 2010. Government as Entrepreneur: Evaluating the Commercialization Success of SBIR Projects. *Research Policy* 39, 5 (2010),

589–601.

[17] Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. In *International Conference on Learning Representations (ICLR)*.

[18] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations (ICLR)*.

[19] Jie Lu, Anjin Liu, Fan Dong, Feng Gu, João Gama, and Guangquan Zhang. 2018. Learning under Concept Drift: A Review. *IEEE Transactions on Knowledge and Data Engineering* 31, 12 (2018), 2346–2363.

[20] Y. Qi, Q. Lu, S. Dou, X. Sun, M. Li, and Y. Li. 2025. Graph Neural Network-Driven Hierarchical Mining for Complex Imbalanced Data. In *2025 8th International Symposium on Big Data and Applied Statistics (ISBDAS)*. IEEE, 320–324. doi:10.1109/ISBDAS64762.2025.11116968

[21] Michael Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling Relational Data with Graph Convolutional Networks. In *European Semantic Web Conference (ESWC)*. Springer, 593–607.

[22] Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S. Yu, and Tianyi Wu. 2011. PathSim: Meta Path-Based Top-K Similarity Search in Heterogeneous Information Networks. *Proceedings of the VLDB Endowment* 4, 11 (2011), 992–1003.

[23] Andrew A. Toole and Dirk Czarnitzki. 2007. Biomedical Academic Entrepreneurship through the SBIR Program. *Journal of Technology Transfer* 32, 4 (2007), 403–417.

[24] U.S. Small Business Administration, Office of Advocacy. 2023. 2023 Small Business Profile. https://advocacy.sba.gov/2023/03/07/2023-small-business-profile/.

[25] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *International Conference on Learning Representations (ICLR)*.

[26] Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S. Yu. 2019. Heterogeneous Graph Attention Network. In *Proceedings of The Web Conference (WWW)*. 2022–2032.

[27] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. 2020. Graph Neural Networks: A Review of Methods and Applications. *AI Open* 1 (2020), 57–81.