Prerequisite Learning through Diversity-driven Selective Attention

Anonymous ACL submission

Abstract

Prerequisite relation extraction aims to identify concept dependencies, which are crucial for curriculum planning and adaptive education. Existing methods struggle with noisy edges, dense graphs, or fail to model diverse concept relations effectively. In this paper, we propose DPPNet, a novel graph-based approach that incorporates a Determinantal Point Process (DPP) to perform diversity-driven neighbor selection, enabling the model to retain informative and structurally diverse relations while discarding redundancy. Our method integrates this pruning mechanism into the learning pipeline and operates in a single pass, leading to a highly efficient and robust model. Empirical results across three benchmark datasets demonstrate that DPPNet outperforms existing state-of-the-art methods across three key dimensions: classification performance (Accuracy and F1-score), memory footprint, and training time. These results highlight DPPNet's effectiveness and scalability, making it a practical choice for real-world educational applications.

1 Introduction

A well-structured curriculum is critical for effective learning, guiding students through concepts in a coherent and pedagogically sound order. While online educational resources such as lectures, tutorials, and videos have become widely accessible, they introduce a key challenge: learners are often left to navigate complex topics without guidance on concept sequencing. Accurately identifying prerequisite relationships among concepts is thus essential for adaptive curriculum design and personalized learning pathways.

Recent work models this task using graph-based representations (Mazumder et al., 2023), where nodes represent documents or concepts and edges denote associations. However, graphs built from large-scale educational corpora are often excessively dense and have many-to-many relationships between documents and concepts, creating highly entangled structures. These dense graphs introduce redundant or noisy edges, increasing computational burden and obscuring the true structure of knowledge dependencies.

To address these challenges, we propose DPPNet, a novel framework for learning prerequisite relations by pruning dense educational graphs using Determinantal Point Processes (DPPs) (Kulesza and Taskar, 2012). DPPNet selects a diverse and informative subset of edges for each node, eliminating redundant or spurious connections without relying on heuristic rules or fixed thresholds. This principled, data-driven pruning improves not only interpretability and accuracy but also reduces memory consumption and training time.

Our main contributions are as follows:

1. We propose a novel graph pruning framework based on Determinantal Point Processes (DPPs), specifically tailored to the structure of educational concept graphs, improving the precision of prerequisite relation extraction.

2. Our approach dynamically selects a diverse and important subset of neighbors for each node, preserving only pedagogically meaningful connections while filtering out redundant or noisy links.

3. We integrate this pruning mechanism into a Graph Attention Network, enabling end-to-end learning over interpretable, semantically-focused, and scalable sparse graphs.

4. The proposed method not only enhances the quality of concept representations but also reduces training time and memory usage significantly, without relying on any additional external information, unlike some state-of-the-art methods.

2 Related Work

The study of concept prerequisite relations (CPRs) has evolved significantly, with advancements in models that capture semantic interconnections between educational concepts and documents. Understanding these relationships is crucial for constructing effective learning paths. Early research focused on quantifying dependencies through textual and structural features, enabling the identification of prerequisite relations (Talukdar and Cohen, 2012). A metric based on hyperlink reference distance was also proposed to assess the closeness of concepts across educational resources (Liang et al., 2015).

Subsequent advancements incorporated probabilistic models and deep learning, such as a model combining topic modeling with Siamese neural networks to improve detection (Roy et al., 2019). Graph-based models, particularly Graph Neural Networks (GNNs), have proven effective in capturing complex relationships. The Variational Graph Autoencoder (VGAE) (Kipf and Welling, 2016) allowed the modeling of latent concept representations, and its extension, the Relational VGAE (R-VGAE) (Li et al., 2020), integrated both concepts and resources. Multi-head attention mechanisms (Zhang et al., 2022) were later incorporated to focus on more informative neighbors, improving accuracy. The graph attention-based model (Mazumder et al., 2023) has also been proposed for concept relation prediction.

Further, supervised methods (Sun et al., 2024a) and evidence-based approaches (Zhang et al., 2024) have also shown promising performance. Weak supervision methods were applied to reduce dependency on labeled data, improving generalization (Zhang et al., 2025a). To further enhance performance, global knowledge graphs and optimization techniques were employed to optimize prerequisite learning (Zhang et al., 2025b), and multiscale GNNs improved link prediction (Zhang et al., 2025c).

Despite advancements, current models tend to treat all graph edges uniformly, which can lead to suboptimal performance. In educational graphs, some connections are more informative than others. Although GNNs perform well, their computational demands are high. Graph pruning strategies (Sun et al., 2024b) that remove less informative edges have been suggested to improve model efficiency, but are time-consuming. The proposed work builds on these insights, aims to optimize the balance between performance, time efficiency, and memory consumption.

3 Proposed Approach

We are given a set of learning resources $\mathcal{R}$ for a topic T and the set of concepts $\mathcal{C}$ present in $\mathcal{R}$. The main objective is to build a model $\mathcal{M}$ that will take a pair of concepts $c_i, c_j \in \mathcal{C}$ and determine if c_i is a prerequisite for c_j. In other words, a learner needs to know the concept c_i before he/she try to know about c_j. This task can naturally be viewed as binary classification. Specifically, if $\mathcal{M}(c_i, c_j) = 1$, then $c_i \in \mathcal{C}$ is the prerequisite concept of $c_j \in \mathcal{C}$, otherwise $c_i \in \mathcal{C}$ is not a prerequisite concept of $c_j \in \mathcal{C}$. We note that, if $\mathcal{M}(c_i, c_j) = 0$, it does not imply that $\mathcal{M}(c_j, c_i) = 0$. Therefore, even though c_i is not a prerequisite for c_j, c_j can still be a prerequisite for c_i.

We assume that the concepts in $\mathcal{R}$ are already known to us. Thus, our task is to uncover the directed dependencies that form meaningful learning sequences. To achieve this, we construct a heterogeneous graph consisting of concept and document nodes, following the methodology proposed in (Mazumder et al., 2023). However, directly operating on this full graph often results in high edge density, especially in large corpora, where redundant or weak connections can negatively impact the model's performance.

As we described in Section 2, several methods have been proposed in the recent past. Unsurprisingly, the methods based on deep neural nets achieve better performance than the classical machine learning models. Furthermore, among the deep learning models, graph neural networks turned out to be the most effective models. However, a major limitation of most of the top-scoring existing methods is that they are not good at generating distinguishable concept representations even for seemingly unconnected concepts, primarily due to the highly connected graph on which they operate. Some methods employ additional annotation (such as the relationship between the documents) to alleviate the problem. However, these methods are not cost-effective.

To address this, we introduce an approach based on a repulsive point process that retains a diverse, informative subset of neighbors for each node from the graph, reducing graph complexity without losing the semantics. We then apply a Graph Attention Network (GAT) to learn node representa-

tions by aggregating information from neighboring nodes. These learned embeddings are then fed into a binary classifier that predicts the relationship between two concepts. Both the GAT and classifier are trained jointly in an end-to-end manner to optimize performance.

In summary, our approach achieves several important goals: (i) enables the graph attention network to generate higher quality concept representations, (ii) reduces order of magnitude training time, (iii) significantly lesser memory footprint, and finally, (iv) does not use any additional inoformation unlike some of the state of the art methods.

Our approach has three major components, namely, *graph construction*, *neighbor selection for each node*, and *Prerequisite Relation Prediction*. We detail below each of these components.

3.1 Graph Construction

We construct a heterogeneous graph $G = (V, E)$ to represent the educational corpus, $V = \mathcal{R} \cap \mathcal{C}$ (the concept nodes and the document nodes).

An edge is created between a document node (d) and a concept node (c) if the concept is present in the document. The weight of the edge is the probability that the frequency f_{dc} of c in d lies in the extreme right tail of its frequency distribution in similar documents (the same measure as in (Mazumder et al., 2023)). An edge between two concepts is present if they have positive pointwise mutual information (PMI) based on a sliding window of length 30, and the weight is the PMI value itself. Finally, for document to document node connections, each document is represented by averaging the vectors of all its concepts. Then, documents are linked using cosine similarity between their concept-based vector representations.

Each node is initialized with a dense embedding. The concept node embeddings are derived from co-occurrence statistics, while document node embeddings are computed as the average of embeddings for the concepts they contain. These embeddings, along with edge weights, serve as inputs for downstream learning in the GAT-based prediction module.

3.2 Neighborhood Refinement

Our objective in this step is to dynamically select a diverse and informative subset of neighbors for each node in a graph, promoting richer and non-redundant neighborhood representations. To achieve this, we adopt a greedy sampling approach inspired by Yao et al. (2016), which efficiently selects a diverse set of neighbors without requiring a fixed k.

Section 3.2.1 provides a brief overview of Determinantal Point Processes (DPPs), while Section 3.2.2 presents the greedy DPP algorithm. In Section 3.2.3, we introduce a novel early filtering mechanism that improves efficiency, and finally, Section 3.2.4 outlines the complete neighbor pruning algorithm.

3.2.1 Background

We leverage the properties of Determinantal Point Processes (DPPs) (Kulesza and Taskar, 2012), which are probabilistic models that favor diverse subsets through negative correlation. Given a kernel matrix $\mathbf{L}$, the probability of selecting a subset $Y \subseteq \mathcal{Y}$ is given by:

$$P(Y) \propto \det(\mathbf{L}_Y) \tag{1}$$

The kernel matrix L captures both the quality and similarity of elements. Each entry is defined as:

$$L_{ij} = q_i \phi_i^\top \phi_j q_j \tag{2}$$

where q_i is a quality score and ϕ_i is a feature vector for the i^{th} item. The inner product $\phi_i^\top \phi_j$ reflects similarity, promoting the selection of dissimilar, high-quality elements. This makes DPPs particularly suitable for edge pruning, where we aim to retain a subset of neighbors that are both relevant and diverse.

While standard DPP sampling favors diversity, its stochastic nature does not ensure optimal or deterministic outcomes. To mitigate this limitation, we employ a greedy approximation of DPPs, as proposed in (Yao et al., 2016), which we detail below.

3.2.2 Greedy DPP Algorithm Overview

The greedy DPP algorithm (Yao et al., 2016) incrementally builds a subset S_i for node i, selecting one element at a time from a candidate pool N_i, which consists of all neighbors of node i. At each iteration, the element with the highest marginal gain in diversity is added:

$$\Delta_s = \text{score}_L(S_i \cup \{s\}) - \text{score}_L(S_i) \tag{3}$$

The diversity score is defined using the log-determinant:

$$\text{score}_L(S_i) = \log\det(L_{S_i}) \tag{4}$$

The process halts when no candidate yields a positive marginal gain. This adaptive strategy retains the diversity-seeking nature of DPPs without needing a fixed subset size.

To ensure that the determinant is positive and the logarithm is well-defined, particularly when the kernel matrix may be singular or nearly singular, we update the score calculation (Equation 4) and add the identity matrix I, thus $\text{score}_L(S_i)$ is given by:

$$\text{score}_L(S_i) = \log\det(L_{S_i} + I) \tag{5}$$

To make this approach more computationally efficient, we introduce an early filtering mechanism that prunes candidates based on their marginal gain, allowing us to reduce the number of candidates to evaluate in each iteration.

3.2.3 Early Filtering Mechanism

The greedy DPP algorithm constructs a diverse subset by iteratively selecting elements that maximize the marginal gain in diversity (Equation 3). However, evaluating Δ_s for all candidates in each iteration becomes computationally expensive for large candidate pools.

To reduce this cost, we introduce an *early filtering mechanism* that prunes candidates with non-positive marginal gain, i.e.,

$$\Delta_s \leq 0 \quad \Rightarrow \quad s \text{ is discarded.}$$

This filtering relies on a fundamental property of DPP. Although the determinant function itself is not submodular, the log-determinant function (Equation 5) is submodular when L is positive semi-definite. This follows from the fact that the log-determinant function behaves like a submodular function as established in (Kulesza and Taskar, 2012). Specifically, for any $S_i \subseteq S_j \subseteq \mathcal{Y}$ and $s \notin S_j$,

$$\begin{aligned}\log\det(L_{S_j\cup\{s\}} + I) - \log\det(L_{S_j} + I) \leq \\ \log\det(L_{S_i\cup\{s\}} + I) - \log\det(L_{S_i} + I)\end{aligned} \tag{6}$$

This inequality expresses the diminishing returns property of the log-determinant function: the marginal contribution of adding an element decreases as the selected set grows. Thus, if an element has zero or negative marginal gain at iteration i, it is guaranteed that it will not contribute positively in subsequent iterations and can therefore be safely pruned.

We define the filtered candidate set as

$$\tilde{N}_i = \{s \in N_i \mid \Delta_s > 0\},$$

from which we select the candidate with the highest marginal gain:

$$s^* = \arg\max_{s\in\tilde{N}_i} \Delta_s.$$

The set is then updated as $S_{i+1} = S_i \cup \{s^*\}$, and the process repeats until $\tilde{N}_i = \emptyset$.

This filtering mechanism significantly reduces the number of determinant evaluations without compromising the diversity-seeking nature of the DPP model. Compared to the baseline greedy DPP algorithm (Yao et al., 2016), our approach improves scalability while maintaining high-quality subset selection.

3.2.4 Neighbor Selection Algorithm

Our algorithm has two main parts: (i) **Kernel Construction:** to build the kernel matrix based on the node's neighborhood and (ii) **Greedy Selection:** a filtering approach to select a diverse subset of neighbors.

This strategy yields adaptive, high-quality neighborhoods per node, without requiring a manually tuned k. The steps are as follows:

Kernel Computation: For node i, identify its neighborhood:

$$N_i = \{j \mid (i,j) \in E\}$$

and construct a kernel $L \in \mathbb{R}^{|N_i|\times|N_i|}$ using:

$$L_{jk} = q_j \cdot \text{sim}(j,k) \cdot q_k \quad \forall j,k \in N_i \tag{7}$$

where $\boldsymbol{q_j}$ is a quality score for the neighbor node j, which is the weight of the edge between the node i and j and $sim(j,k)$ is the similarity between j and k. The kernel inherently favors sets of nodes that are individually relevant to i, but dissimilar to one another, helping avoid redundancy in the selected neighborhood.

Greedy Selection: Once the kernel matrix L is constructed (Equation 7) for a given node i, we apply the greedy selection algorithm to select a diverse subset of neighbors from the candidate pool N_i.

At each iteration, the algorithm evaluates the marginal gain Δ_s using the log-determinant score (Equation 5), and applies the early filtering mechanism (Section 3.2.3) to discard candidates with non-positive gain.

The detailed procedure for neighbor selection of each node i in graph $G = (V, E)$ using kernel L is outlined in the algorithm 1.

Algorithm 1 Greedy Selection Algorithm

Input: Neighbor N_i, Kernel L & threshold ϵ.
Output: Selected Neighbor N_i'

$S_i \leftarrow \emptyset$, $C \leftarrow N_i$
while $C \neq \emptyset$ **do**
  for all $s \in C$ **do**
    Compute Δ_s using Eq. (3)
  end for
  $s^* \leftarrow \arg\max_{s \in C} \Delta_s$
  if $\Delta_{s^*} \leq \epsilon$ **then**
    break
  end if
  Add s^* to S_i, remove s^* from C
  Remove all $s \in C$ where $\Delta_s \leq 0$
end while
$N_i' \leftarrow S$
return N_i'

Although the input graph is undirected, we perform neighbor pruning in an asymmetric manner. For each node i, we apply the selection algorithm (Alg. 1) to choose a diverse subset of neighbors $N_i' \subseteq N_i$. The connections from i to nodes not in N_i' are removed only with respect to i. That is, the edge (i, j) is removed from i's perspective if $j \notin N_i'$, but it may still exist from j's perspective if $i \in N_j'$. This results in an effectively asymmetric neighborhood structure, even though the original graph is undirected.

3.3 Concept Relation Prediction

After the DPP-based pruning phase (Section 3.2), we obtain a sparsified concept graph $G' = (V, E')$ that retains high-quality and diverse connections. We now utilize this graph to learn task-specific node representations and identify prerequisite relationships between concepts. For this purpose, we adopt a two-stage neural framework introduced in prior work (Mazumder et al., 2023), comprising a Graph Attention Network (GAT) for node encoding and a pairwise classifier for relation prediction.

3.3.1 Node Representation

The first stage of the model employs a Graph Attention Network (Velickovic et al., 2018) to encode each node based on its local neighborhood.

The input consists of node features derived from embeddings for both concept and document nodes. Each node i is associated with a feature vector $\mathbf{v}_i \in \mathbb{R}^F$, which is first linearly transformed into a higher-level space via a shared weight matrix $\Theta \in \mathbb{R}^{F' \times F}$. To capture contextual relevance, an attention score is computed between each node i and its neighbors $j \in \mathcal{N}_i$. The unnormalized attention coefficient c_{ij} is computed using a single-layer feedforward network with LeakyReLU activation:

$$c_{ij} = \text{LeakyReLU}\left(\mathbf{a}^\top [\Theta \mathbf{v}_i \parallel \Theta \mathbf{v}_j]\right)$$

where $\mathbf{a} \in \mathbb{R}^{2F'}$ is a learnable attention weight vector and $\parallel$ denotes vector concatenation. These coefficients are then normalized across the neighborhood using softmax:

$$\alpha_{ij} = \frac{\exp(c_{ij})}{\sum_{k \in \mathcal{N}_i} \exp(c_{ik})}$$

In the case of graphs with edge features, it can be incorporated by extending the attention computation to include transformed edge embeddings:

$$c_{ij} = \text{LeakyReLU}\left(\mathbf{a}^\top [\Theta \mathbf{v}_i \parallel \Theta \mathbf{v}_j \parallel \Theta_e \mathbf{e}_{ij}]\right)$$

where $\mathbf{e}_{ij}$ denotes the edge feature. The final representation for each node i is then a weighted aggregation over its neighbors:

$$\mathbf{v}_i' = \sigma\left(\sum_{j \in \mathcal{N}_i} \alpha_{ij} \cdot \Theta \mathbf{v}_j\right)$$

This results in task-aware, neighborhood-sensitive embeddings that reflect both local structure and semantic importance.

3.3.2 Relation Prediction

The second stage of the model is a pairwise prediction module. Given the learned embeddings $\mathbf{v}_i'$ and $\mathbf{v}_j'$ for a candidate concept pair (C_i, C_j), the model predicts whether $C_i \rightarrow C_j$ holds. Each embedding is first passed through a feedforward network with shared weights:

$$\mathbf{h}_i = \text{ReLU}(W_s \mathbf{v}_i' + \mathbf{b}_s), \;\; \mathbf{h}_j = \text{ReLU}(W_s \mathbf{v}_j' + \mathbf{b}_s)$$

The final relation score is computed from a joint representation formed by combining the two hidden vectors:

$$\mathbf{x}_{ij} = [\mathbf{h}_i; \mathbf{h}_j; \mathbf{h}_i - \mathbf{h}_j; \mathbf{h}_i \odot \mathbf{h}_j]$$

$$p(C_i \rightarrow C_j) = \sigma(W^\top \mathbf{x}_{ij} + b)$$

where $\odot$ denotes element-wise multiplication and σ is the sigmoid function.

The model is trained using binary cross-entropy loss:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{|D|} \sum_{(i,j,y_{ij}) \in D} [y_{ij} \log p_{ij} + (1 - y_{ij}) \log(1 - p_{ij})] \quad (8)$$

where $y_{ij} \in \{0, 1\}$ is the ground truth label indicating whether C_i is a prerequisite for C_j.

4 Experimental Setup

This section presents the experimental results obtained on three different datasets. We evaluate the performance of our proposed approach, DPPNet, DPP Pruned Graph for Attention Prerequisite Network, in comparison with several other graph neural network-based models for prerequisite learning. For assessment, we employ standard evaluation metrics, including precision, recall, F1 score, and accuracy.

Table 1: Dataset Statistics

Dataset	$\lvert\mathcal{D}\rvert$	$\lvert\mathcal{C}\rvert$	$\lvert C_{preq}\rvert$
Lecture Bank	277	320	821
MOOC	382	406	4332
University Course	654	407	4347

4.1 Datasets

We perform experiments on three publicly available benchmark educational datasets. Table 1 displays the statistics for all three datasets. In this table, the column $|\mathcal{D}|$ represents the total number of documents, $|\mathcal{C}|$ indicates the total number of concepts and $|\mathcal{C}_{Preq}|$ refers to the total count of concept prerequisite relationships. A detailed description of each dataset is provided below.

- **Massive Open Online Course (MOOC)**[1]: This dataset is sourced from the Massive Open Online Course (MOOC), as used in (Pan et al., 2017). It includes concepts related to Computer science and comprises 406 concepts from various university-level courses. Each course is accompanied by multiple video lectures, along with subtitles, where each subtitle represents a distinct document.

- **Lecture Bank (LB)**[2] **:** This dataset (Li et al., 2019) contains English lecture files from 60 courses covering 5 different domains, including Natural Language Processing (NLP), Machine Learning (ML), Artificial Intelligence (AI), deep learning (DL), and information retrieval (IR).

- **University Course (UC)**[3]**:** Introduced by (Liang et al., 2017), the university course dataset contains course descriptions from various university courses. These courses include subjects like Algorithm Design, Computer Graphics, Graph Theory, and Neural Networks from the domain of computer science.

4.2 Baselines

We compare our proposed method, DPPNet, with eleven state-of-the-art models for concept prerequisite relation prediction. These models include RefD (Liang et al., 2015), M3 (Miaschi et al., 2019), GAE and VGAE (Li et al., 2019), PREREQ (Roy et al., 2019), R-VGAE(T) and R-VGAE(P) (Li et al., 2020), MHAVGAE (Zhang et al., 2022), HGAPNet (Mazumder et al., 2023), LCPRE (Sun et al., 2024b) and GKROM (Zhang et al., 2025b). These baselines represent key advancements in prerequisite relation extraction, specifically graph-based neural networks, multi-objective knowledge optimization, and learning-path-based sparse graph approaches. Each baseline brings a unique perspective to the task, and comparing them allows us to showcase the advantages of our proposed approach in terms of both performance and computational efficiency. Each method is trained with the same 8:1:1 ratio of data for fair comparison.

4.3 Implementation Details

We use the concept prerequisite relations given by (Zhang et al., 2025b) and adopt an 8:1:1 ratio to divide the dataset into training, validation, and test sets. The model is trained for 500 epochs with a batch size of 4 using the Adam optimizer, with binary cross-entropy employed as the loss function. Consistent with the configuration in (Mazumder et al., 2023), our architecture includes two graph attention layers: the first with 128 hidden units and the second with 512. The prediction component is

[1] http://keg.cs.tsinghua.edu.cn/jietang/software/acl17-prerequisite-relation.rar

[2] https://github.com/Yale-LILY/LectureBank

[3] https://github.com/suderoy/PREREQ-IAAI-19/

a feed-forward layer that maps a 512-dimensional input to a 64-dimensional output vector. All experiments are conducted on a system equipped with an NVIDIA A100 GPU with 80 GB of memory, an Intel Xeon Gold 6330 CPU running at 2.00 GHz, and 376 GB of RAM.

5 Results

We evaluate our proposed model, DPPNet, on three widely used educational datasets. The evaluation covers two key aspects: the model's ability to accurately extract prerequisite relations and its computational efficiency in terms of memory usage and training time.

5.1 Performance on Prerequisite Relation Extraction

The evaluation, based on F1-score and accuracy, is summarized in Table 2. DPPNet outperforms all baselines, achieving the highest scores on all three datasets. Among the baselines, HGAPNet ranks second, showing solid generalization, while models like LCPRE and GKROM perform well on specific datasets but lack consistency. Older methods such as PREREQ, GAE, and VGAE struggle with complex prerequisite relationships, and models like MHAVGAGE and R-VGAE(P) show only modest improvements. Notably, while LCPRE excels on LectureBank, it does not generalize across other datasets.

These results demonstrate that DPPNet excels in both accuracy and consistency, offering a strong balance between precision and recall as indicated by its superior F1-scores

5.2 Computational Efficiency

In real-world educational systems, strong model performance must be balanced with computational efficiency for scalability and usability. To evaluate this, we compare the memory usage, training time, and edge sparsification of our proposed model (DPPNet) against leading baselines—GKROM, LCPRE, and HGAPNet—across three benchmark datasets. These baselines were chosen due to their competitive accuracy and F1-score (Table 2), while other methods were excluded for their lower performance and practical viability. The following sections delve into a comprehensive analysis of each aspect.

5.2.1 Memory Utilization

Table 3 shows the memory consumption (in GB) across all datasets. DPPNet exhibits the lowest memory usage, outperforming LCPRE, which also uses graph sparsification. DPPNet's pruning strategy, based on Determinantal Point Processes, efficiently removes redundant edges, resulting in more compact graphs. In contrast, HGAPNet and GKROM use dense graphs with complex relational modeling, leading to significantly higher memory consumption. These results highlight DPPNet's scalability, particularly in low-resource settings or large educational datasets.

5.2.2 Training Time Comparsion

Beyond memory savings, DPPNet also delivers substantial gains in training speed, as shown in Table 4. Across all datasets, it consistently trains in less than one-third of the time required by its closest competitors. In the baselines, based on the data, it can be seen that HPANet performs better than the other two baselines.

It is particularly notable that LCPRE, despite using sparsification to reduce memory, still suffers from longer training times. This likely stems from its added temporal modeling and path-based reasoning, which introduce complexity during training. DPPNet, in contrast, uses a *single-shot, diversity-driven pruning mechanism*, reducing not only the graph size but also the computation needed for each learning iteration.

5.2.3 Edge Sparsification

An essential feature of DPPNet is its ability to significantly reduce graph density while preserving task-relevant information. Figure 1 compares edge counts (log scale) across methods on the MOOC dataset, categorized by edge types: CC (Concept-Concept), DD (Document-Document), and DC (Document-Concept). We observe the same pattern in the two datasets as well.

HPAGNet and GKROM use the full graph, resulting in high edge counts and computational cost. LCPRE performs moderate pruning but retains substantial edge density. In contrast, DPPNet achieves over 98% edge reduction, maintaining strong classification performance. The significant reduction in DD (Document-Document) edges by DPPNet and LCPRE suggests that document-level connections contribute minimally to concept prerequisite classification and may introduce more noise than value.

Table 2: Performance Comparison. Best results are bolded, and runner-ups are underlined.

Method	University Course		LectureBank		MOOC	
	ACC	F1	ACC	F1	ACC	F1
RefD	0.762	0.711	0.739	0.757	0.818	0.714
M3	0.825	0.821	0.794	0.786	0.781	0.690
GAE	0.664	0.663	0.687	0.687	0.671	0.670
VGAE	0.694	0.698	0.714	0.711	0.675	0.676
PREREQ	0.543	0.587	0.510	0.556	0.512	0.582
R-VGAE(T)	0.685	0.682	0.666	0.644	0.593	0.535
R-VGAE(P)	0.737	0.720	0.702	0.661	0.703	0.663
MHAVGAGE	0.788	0.795	0.726	0.740	0.748	0.764
HGAPNet	0.871	0.875	0.787	0.780	0.882	0.888
LCPRE	0.820	0.829	0.830	0.846	0.845	0.852
GKROM	0.870	0.874	0.823	0.820	0.863	0.869
DPPNet (Ours)	**0.886**	**0.891**	**0.860**	**0.852**	**0.889**	**0.895**

Table 3: Memory Usage Comparison (in GB). Best results are bolded, and runner-ups are underlined.

Method	LB	MOOC	UC
HGAPNet	3.11	4.97	8.59
LCPRE	0.78	0.78	0.78
GKROM	3.11	4.97	8.58
DPPNet (Ours)	**0.65**	**0.65**	**0.67**

Table 4: Computational Time Comparison (in Hours).

Method	LB	MOOC	UC
HGAPNet	0.65	4.90	8.35
LCPRE	1.47	8.73	8.62
GKROM	0.81	5.56	8.82
DPPNet (Ours)	**0.26**	**1.34**	**1.27**

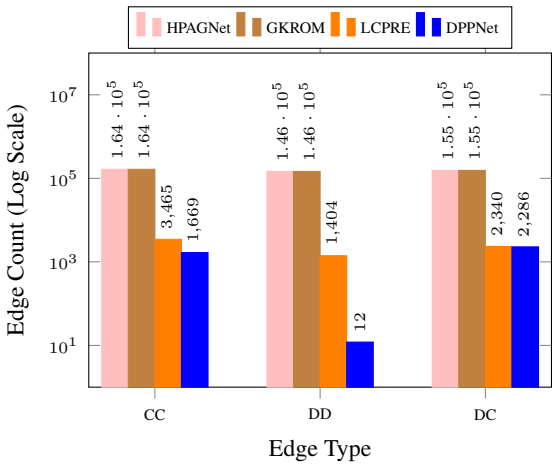

Figure 1: Edge Count Comparison on MOOC Dataset.

6 Conclusion

In this paper, we introduced DPPNet, a novel method for concept prerequisite relation extraction that utilizes Determinantal Point Process (DPP)-based graph pruning. Our approach addresses the challenge of balancing prediction accuracy and computational efficiency by selectively retaining the most informative edges, thus reducing graph size without sacrificing effectiveness. DPPNet's lightweight pruning mechanism not only enhances memory usage and reduces training time but also improves generalization by eliminating noisy or redundant connections. This contrasts with the common assumption that dense graphs are more expressive, demonstrating that sparse graphs, when carefully constructed, can achieve comparable or superior results.

Our experimental results confirm that DPPNet outperforms other state-of-the-art methods, including approaches that retain all edges (HGAPNet), those based on a pruned structure (LCPRE), and even models that incorporate additional external knowledge (GKROM). DPPNet's ability to dynamically select the number of edges to prune, without requiring predefined inputs, further highlights its adaptability and scalability, making it a promising solution for large-scale educational applications. These findings not only set a new standard in concept prerequisite relation extraction but also pave the way for more resource-efficient and interpretable models in educational content design.

Limitations

While DPPNet represents a significant advancement in concept prerequisite relation extraction, there are a few limitations that need to be considered for future improvements.

- Dependency on Graph Quality: The effectiveness of the DPP-based pruning approach is heavily reliant on the quality of the input graph. If the graph construction is flawed or incomplete, the pruning process may inadvertently remove important connections, potentially reducing the model's accuracy. Thus, ensuring high-quality graph construction remains a key challenge for improving performance.

- Scalability for Extremely Large Datasets: While DPPNet demonstrates strong performance on large-scale educational datasets, its scalability may face challenges when dealing with extremely large or highly complex graphs. The sheer volume of data in such cases could result in increased computation times. Although pruning techniques help reduce model complexity, the process of selecting edges from massive graph structures may still impose significant computational overhead, potentially limiting efficiency for very large datasets.

- Domain-Specific Adaptation: The model's performance might vary across different domains or educational contexts. DPPNet has been evaluated on a few specific datasets, and while it has shown strong performance, its generalization to other fields with significantly different learning structures or concept relationships remains an open question. Further research into domain adaptation techniques could enhance its applicability across diverse educational domains.

Despite these limitations, DPPNet provides a strong foundation for future research and development in the field of concept prerequisite relation extraction. Addressing these challenges in future work can pave the way for even more robust, scalable, and interpretable models.

References

Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. In *Proceedings of the 2016 International Conference on Neural Information Processing Systems*, pages 2148–2156. Curran Associates, Inc.

Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5.

Irene Li, Alexander R. Fabbri, Swapnil Hingmire, and Dragomir R. Radev. 2020. R-VGAE: relational-variational graph autoencoder for unsupervised prerequisite chain learning. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 1147–1157. International Committee on Computational Linguistics.

Irene Li, Alexander R. Fabbri, Robert R. Tung, and Dragomir R. Radev. 2019. What should i learn first: introducing lecturebank for nlp education and prerequisite chain learning. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence*.

Chen Liang, Zhaohui Wu, Wenyi Huang, and C. Lee Giles. 2015. Measuring prerequisite relations among concepts. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1668–1674.

Chen Liang, Jianbo Ye, Zhaohui Wu, Bart Pursel, and C. Giles. 2017. Recovering concept prerequisite relations from university course dependencies. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31.

Debjani Mazumder, Jiaul H. Paik, and Anupam Basu. 2023. A graph neural network model for concept prerequisite relation extraction. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 1787–1796.

Alessio Miaschi, Chiara Alzetta, Franco Alberto Cardillo, and Felice Dell'Orletta. 2019. Linguistically-driven strategy for concept prerequisites learning on Italian. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 285–295.

Liangming Pan, Chengjiang Li, Juanzi Li, and Jie Tang. 2017. Prerequisite relation learning for concepts in MOOCs. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1447–1456.

Sudeshna Roy, Meghana Madhyastha, Sheril Lawrence, and Vaibhav Rajan. 2019. Inferring concept prerequisite relations from online educational resources. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):9589–9594.

Jingwen Sun, Yu He, Yiyu Xu, Jingwei Sun, and Guangzhong Sun. 2024a. A learning-path based supervised method for concept prerequisite relations extraction in educational data. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 2168–2177.

Jingwen Sun, Yu He, Yiyu Xu, Jingwei Sun, and Guangzhong Sun. 2024b. A learning-path based supervised method for concept prerequisite relations extraction in educational data. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 2168–2177.

Partha P Talukdar and William W Cohen. 2012. Learning to predict prerequisite relations from text. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing*, pages 1437–1446. Association for Computational Linguistics.

Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *6th International Conference on Learning Representations, ICLR 2018*.

Jin-ge Yao, Feifan Fan, Wayne Xin Zhao, Xiaojun Wan, Edward Chang, and Jianguo Xiao. 2016. Tweet timeline generation with determinantal point processes. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, page 3080–3086.

JiaYu Zhang, XiaoYan Zhang, XiaoFeng Du, and TianBo Lu. 2024. Ebcpl: A novel evidence-based method for concept prerequisite relation learning. In *International Conference on Database Systems for Advanced Applications*, pages 390–405. Springer.

Juntao Zhang, Nanzhou Lin, Xuelong Zhang, Wei Song, Xiandi Yang, and Zhiyong Peng. 2022. Learning concept prerequisite relations from educational data via multi-head attention variational graph auto-encoders. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 1377–1385.

Miao Zhang, Jiawei Wang, Kui Xiao, Zhifang Huang, Zhifei Li, and Yan Zhang. 2025a. Enhancing weak supervision for concept prerequisite relation learning. *IEEE Transactions on Big Data*.

Miao Zhang, Jiawei Wang, Kui Xiao, Shihui Wang, Yan Zhang, Hao Chen, and Zhifei Li. 2025b. Learning concept prerequisite relation via global knowledge relation optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):1638–1646.

Yupei Zhang, Xiran Qu, Shuhui Liu, Yan Pang, and Xuequn Shang. 2025c. Multiscale weisfeiler-leman directed graph neural networks for prerequisite-link prediction. *IEEE Transactions on Knowledge and Data Engineering*.