COL761: Data Mining
Assignment 1: Frequent Subgraph Mining and Indexing
Question 3: Discriminative Subgraph Identification

Team Members:
Anant Gupta (2025AIB2556)
Vedant Gupta (2025AIB2559)
Devanshu Patel (2025AIB2567)

Indian Institute of Technology Delhi
February 3, 2026

1 Methodology: Discriminative Subgraph Identification

Our approach utilizes an **Ultra-Optimized Bit-Vector Orthogonalization Strategy** to balance the two competing requirements of graph indexing: Coverage (ensuring no query results in a full-database scan) and Precision (minimizing the candidate set size $|C_q|$).

1.1 Greedy Bit-Vector Orthogonalization (Optimization)

Originally, simple frequency-based selection led to redundant features for topologically similar structures. We transitioned to a variance-diversity approach:

- **Variance Pass**: We rank subgraphs by their distance to the 50% frequency mark (maximum Entropy). A feature found in 50% of the database provides the maximum information gain for binary pruning.
- **Orthogonalization Pass**: From the top candidate pool, we greedily select 50 features that maximize the unique variance. For each new feature, we penalize it based on its correlation (bit-vector overlap) with the already selected set.

1.2 FSM Integration: Low-Support Gaston Mining

Instead of relying on high-support heuristics, we push the mining boundary to discover rare but highly discriminative skeletons:

- **Deep Mining**: We run Gaston at very low support levels (1.5% for NCI and 4% for Mutagenicity) to extract a massive pool of structural candidates.
- **Label Mapping**: Our implementation uses a dynamic label mapping system to handle alphanumeric labels transparently, ensuring compatibility across diverse chemical and biological domains.

2 Justification: Why Gaston?

We selected **Gaston** as the core engine for this indexing task due to several critical factors:

- **Performance**: Gaston efficiently mines both trees and cyclic subgraphs. For 40,000 graphs at 1.5% support, it completes in seconds, providing a superior feature pool compared to path-only miners.
- **Completeness**: Unlike custom path miners, Gaston ensures we don't miss complex ring-based structures which are essential for precision in chemical datasets.

3 Database and Domain Analysis

Analysis of NCI-H23 and Mutagenicity shows that these datasets are structurally diverse. Mutagenicity is heavily dominated by nitrogen-based aromatic rings, while NCI-H23 exhibits a wider range of organic skeletons. By using bit-vector analysis, we confirmed that atomic labels alone are insufficient for pruning, requiring the indexing of structural "skeletons" to achieve high filtering rates.

4 Experimental Results and Evaluation

To evaluate the effectiveness, we implemented an assessment based on the grading formula: $S_q = |R_q|/|C_q|$. Our ultra-optimized algorithm achieved record-breaking pruning performance.

Dataset	Total DB Size	Average $\|C_q\|$	Average S_q Score	Space Pruned
Mutagenicity	4,337	**1119.32**	**0.248**	**74.2%**
NCI-H23	40,353	**8313.66**	**0.186**	**79.4%**

Table 1: Overall Performance Metrics (k=50)

4.1 Execution Efficiency

The entire pipeline is designed for high-performance execution. Feature identification for NCI-H23 (40k graphs) completes in approximately 20 minutes, which is well within the 1-hour project constraint.

5 References and Attributions

- **Methodology & Code Implementation**: Developed by the team using the Gaston FSM engine and bit-vector diversity selection.
- **Strategic Refinement**: Technical logic for the orthogonalization and label mapping was refined in collaboration with **Gemini**.
- **Report Compilation**: The structure, analysis, and LaTeX formatting were prepared with assistance from **ChatGPT**.