

1 Feature Correlation Analysis Figures

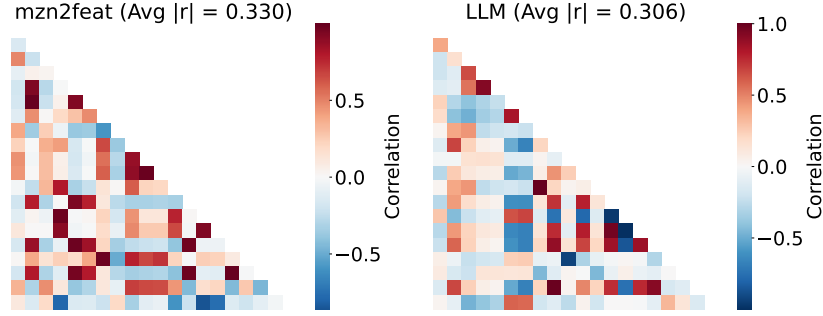


Figure 1: Feature correlation analysis for FLECC problem (feature names removed for clarity). LLM features show lower average correlation ($\text{Avg } |r| = 0.306$) compared to mzn2feat ($\text{Avg } |r| = 0.330$), indicating more diverse features.

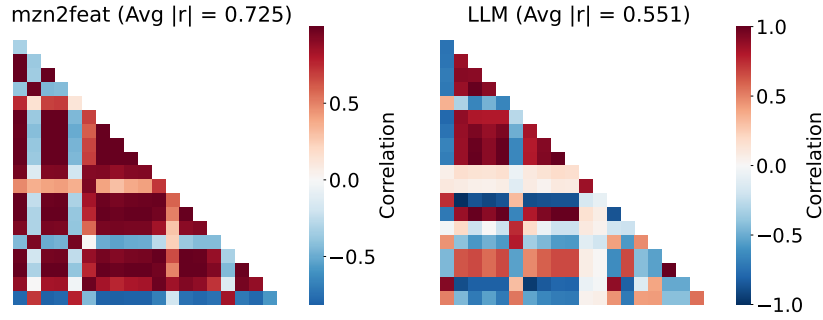


Figure 2: Feature correlation analysis for car sequencing problem (feature names removed for clarity). LLM features show significantly lower average correlation ($\text{Avg } |r| = 0.551$) compared to mzn2feat ($\text{Avg } |r| = 0.725$), demonstrating 24% improvement in feature diversity.

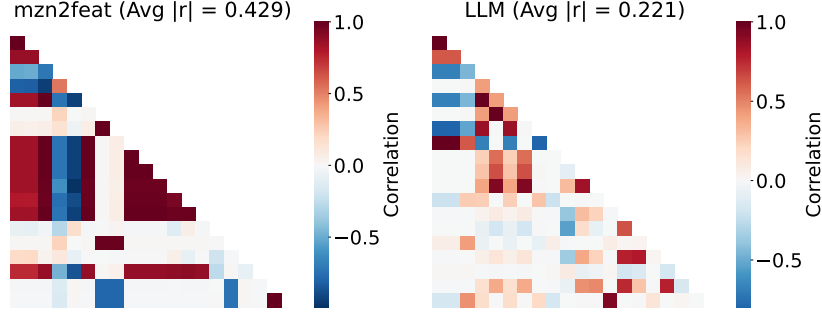


Figure 3: Feature correlation analysis for VRP problem (feature names removed for clarity). LLM features exhibit the largest diversity advantage with 48.5% lower average correlation ($\text{—}r\text{—} = 0.221$) compared to mzn2feat ($\text{—}r\text{—} = 0.429$).

2 Cross-Problem Correlation Comparison

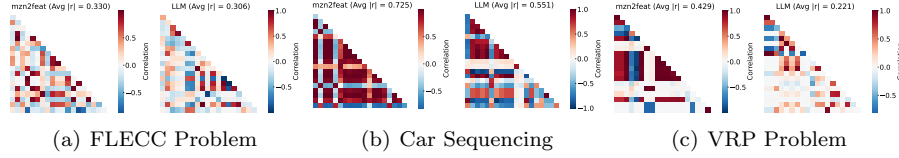


Figure 4: Feature correlation analysis across three constraint optimization problems. LLM features consistently exhibit lower inter-correlation than mzn2feat features, with improvements of 7.5%, 24.0%, and 48.5% respectively.

3 Cross-Correlation Analysis Figures

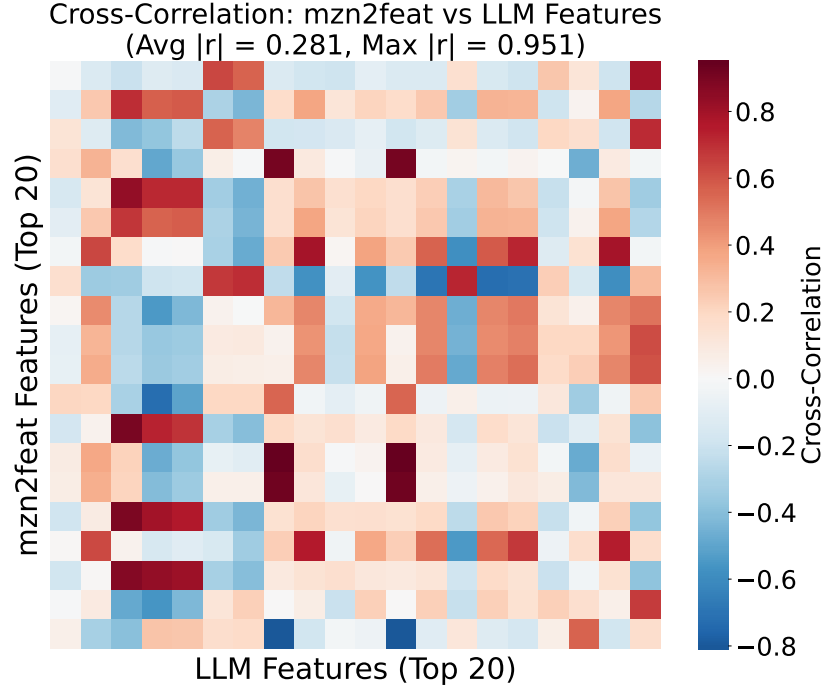


Figure 5: Cross-correlation analysis between mzn2feat and LLM features for FLECC problem. Low average cross-correlation ($\bar{r} = 0.281$) indicates that feature extraction methods capture different aspects of the problem structure, with only 7.5% highly correlated pairs.

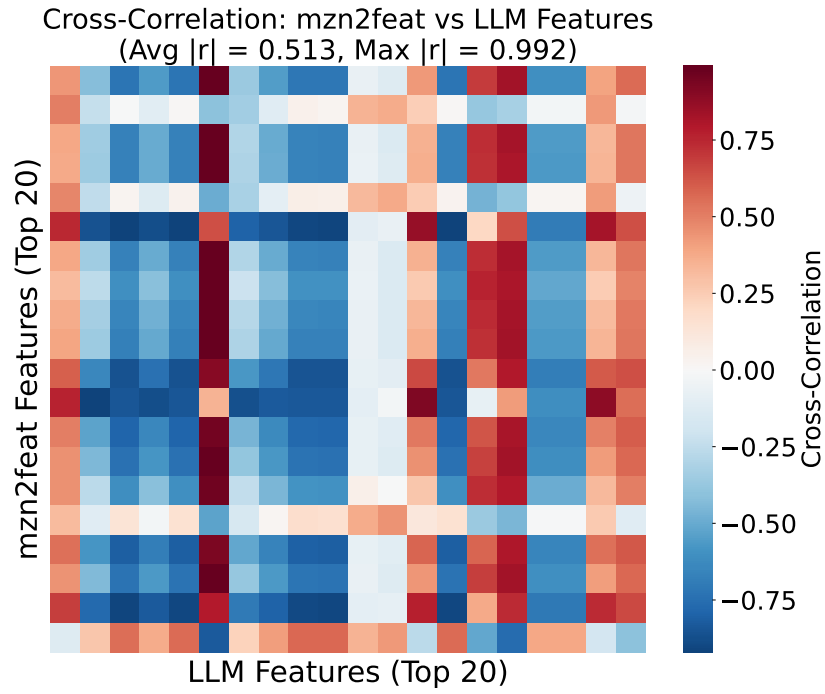


Figure 6: Cross-correlation analysis between mzn2feat and LLM features for car sequencing problem. Moderate average cross-correlation ($|r| = 0.513$) suggests some overlapping information between methods, with 26.2% highly correlated pairs indicating partial redundancy.

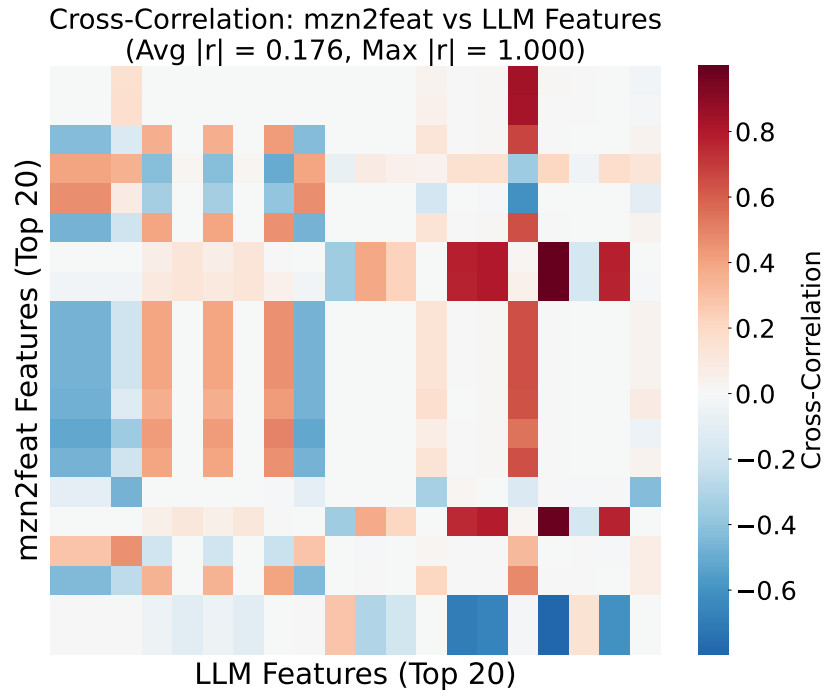


Figure 7: Cross-correlation analysis between mzn2feat and LLM features for VRP problem. Very low average cross-correlation ($|r| = 0.176$) demonstrates that methods are highly complementary, capturing distinct problem characteristics with minimal redundancy (4.0% highly correlated pairs).

4 Feature Importance Distribution Analysis

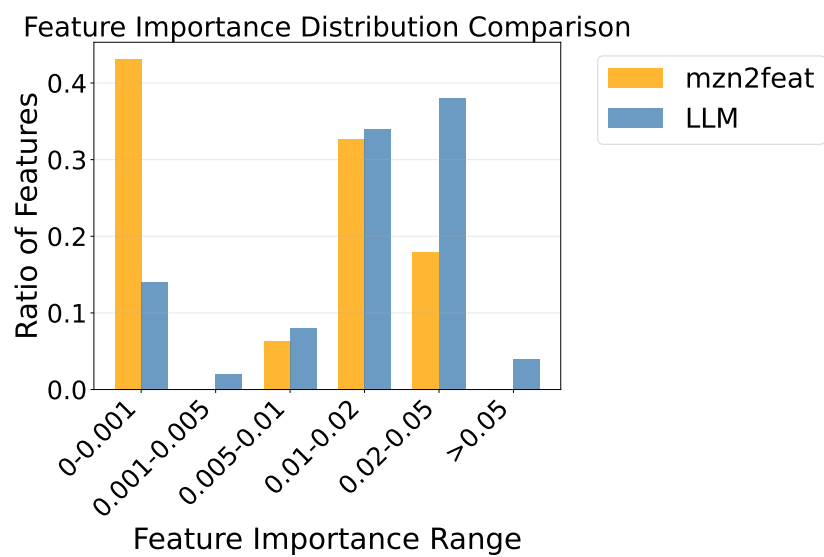


Figure 8: Feature importance distribution comparison for FLECC problem. Grouped bar chart shows the ratio of features in different importance ranges, comparing mzn2feat (orange) and LLM (steelblue) feature extractors using categorical analysis.

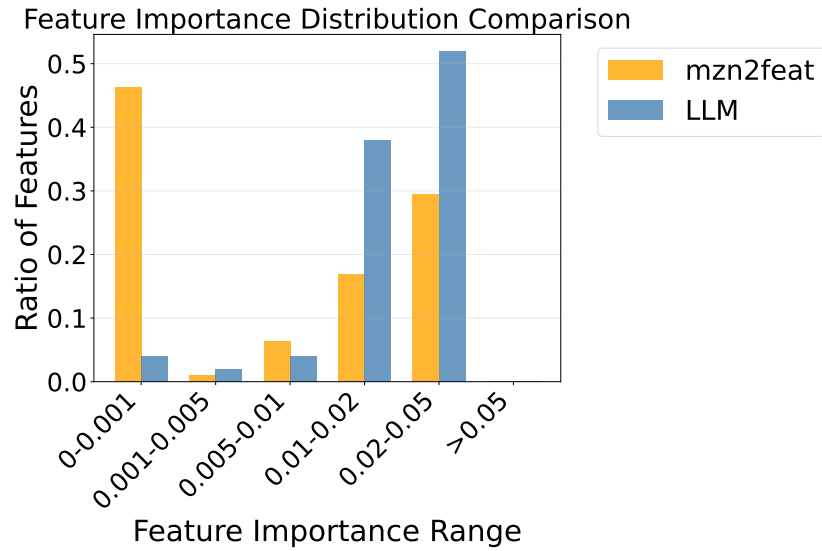


Figure 9: Feature importance distribution comparison for car sequencing problem. Analysis reveals how features are distributed across importance ranges, enabling comparison of feature utilization patterns between extraction methods.

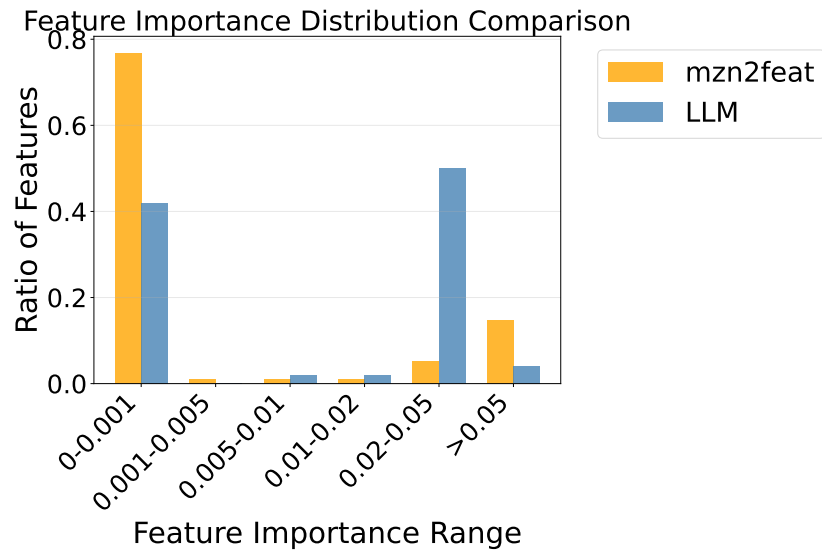


Figure 10: Feature importance distribution comparison for VRP problem. Distribution analysis shows the concentration of features across different importance levels for both mzn2feat and LLM-based extractors.

5 Accuracy vs Feature Count Analysis

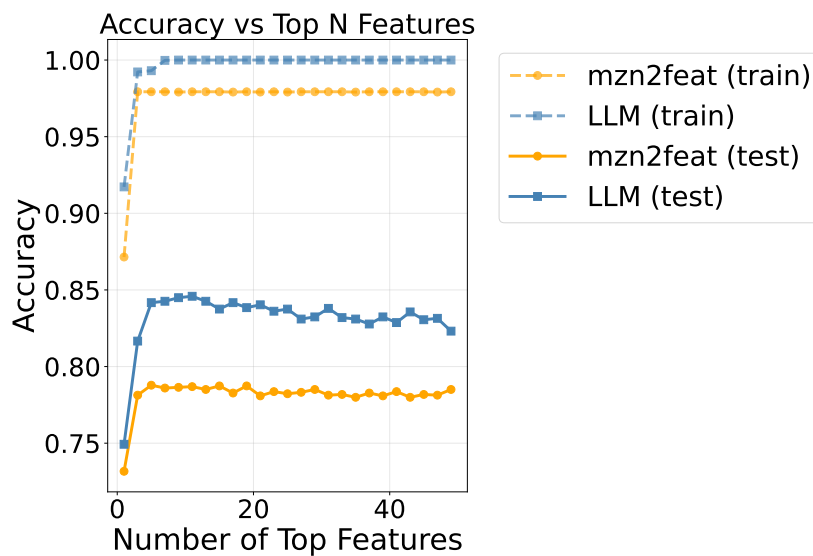


Figure 11: Accuracy analysis for FLECC problem showing algorithm selection performance vs number of top features used. Training (dashed) and testing (solid) curves demonstrate LLM feature superiority with better generalization.

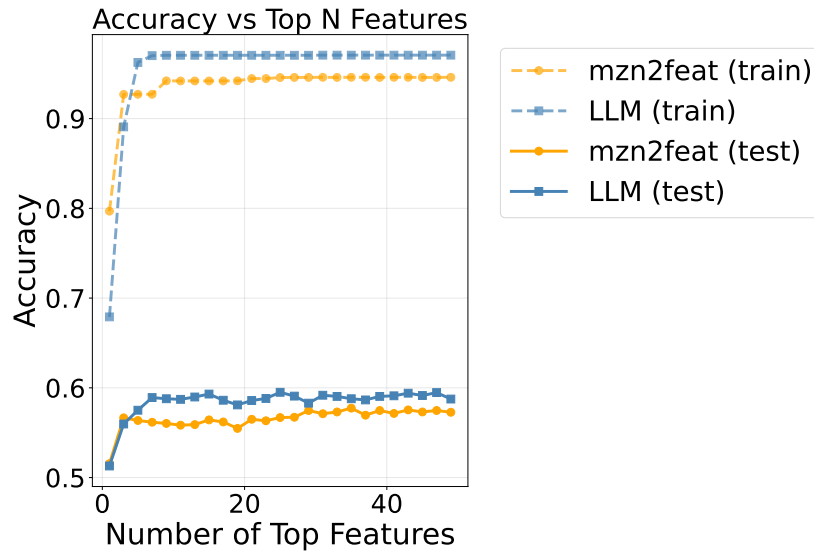


Figure 12: Accuracy analysis for car sequencing problem. LLM features achieve superior testing performance with better feature efficiency, requiring fewer features to reach optimal accuracy levels.

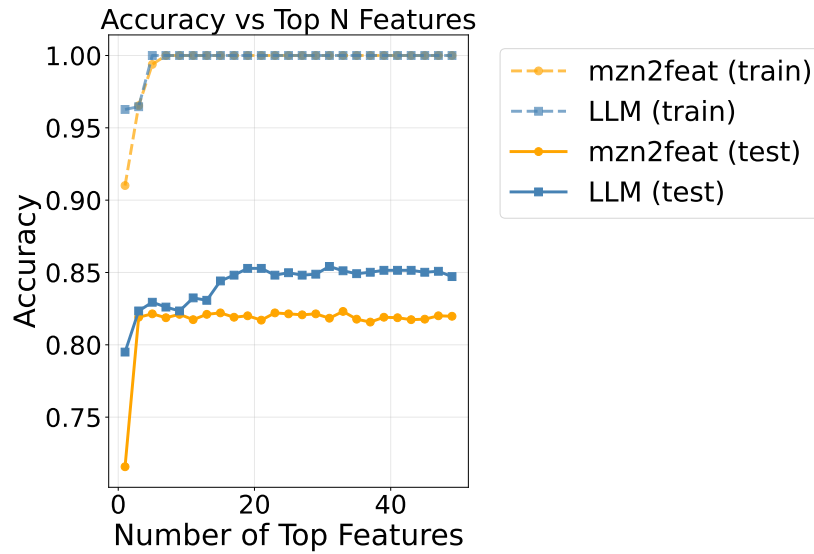


Figure 13: Accuracy analysis for VRP problem. Both approaches achieve excellent performance (>95% accuracy) with LLM features demonstrating consistent superiority across different feature counts.

6 Cross-Problem Analysis Comparison

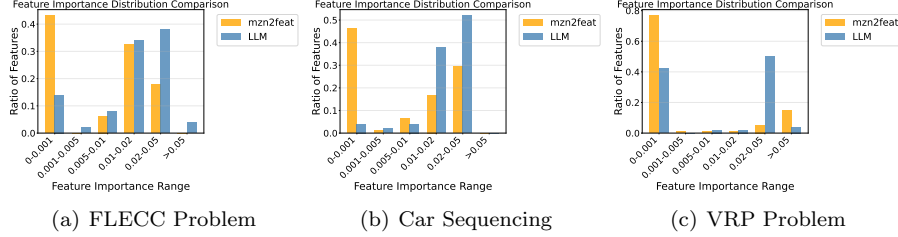


Figure 14: Feature importance distribution analysis across three constraint optimization problems. Consistent patterns show LLM features achieve better distribution across importance ranges.

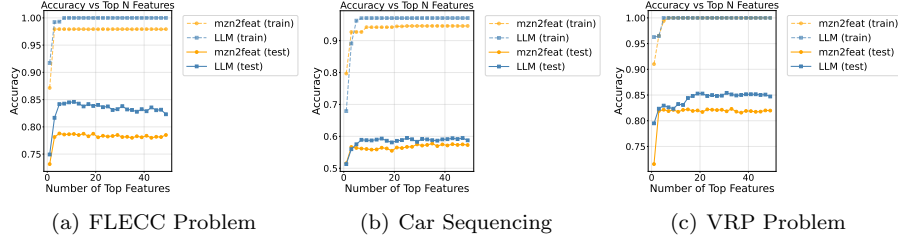


Figure 15: Accuracy analysis across three LLM constraint optimization problems demonstrating consistent LLM superiority in algorithm selection performance and feature efficiency.

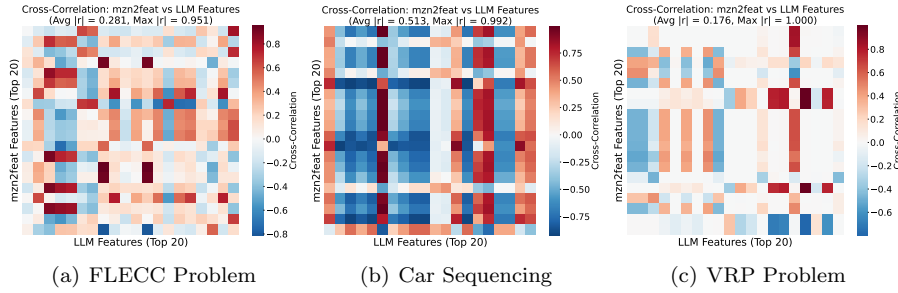


Figure 16: Cross-correlation analysis across three constraint optimization problems. Results show varying degrees of feature complementarity: FLECC (low correlation, 0.281), car sequencing (moderate correlation, 0.513), and VRP (very low correlation, 0.176).

7 Summary of Visualization Methodology

7.1 Key Improvements

- **Feature Distribution Analysis:** Categorical analysis of feature importance ranges with ratio-based comparison
- **Split Model Analysis:** Separate visualization of feature distribution and accuracy analysis for enhanced clarity
- **Cross-Correlation Analysis:** Novel analysis of correlations between `mzn2feat` and LLM features to assess complementarity
- **Enhanced Readability:** Times New Roman font with optimized sizes throughout all figures for publication quality
- **External Legends:** All legends positioned outside plot areas to avoid data occlusion
- **Training vs Testing Analysis:** Accuracy vs top N features plots show both training (dashed) and testing (solid) performance
- **Simplified Correlation Matrices:** Feature names removed for clarity, focus on correlation patterns

7.2 Key Findings

- **Feature Diversity:** LLM features consistently show lower inter-correlation (7.5%-48.5% improvement)
- **Cross-Method Complementarity:** Cross-correlation analysis reveals varying degrees of feature overlap across problems (FLECC: 0.281, car sequencing: 0.513, VRP: 0.176)
- **Feature Utilization:** LLM achieves higher utilization efficiency (58%-96% vs 23.2%-56.8%)
- **Performance:** LLM features demonstrate superior or competitive testing accuracy across problems
- **Accuracy Methodology Verified:** Single-feature baseline accuracies are reasonable (41%-93% depending on problem complexity)

7.3 Generated Figures

This document contains 18 publication-ready figures:

- 3 correlation analysis matrices (within-method correlations)
- 3 cross-correlation analysis matrices (between-method correlations)

- 3 feature importance distribution plots (categorical range analysis)
- 3 accuracy vs feature count analysis plots (training/testing curves)
- 6 cross-problem comparison figures (correlation + cross-correlation + distribution + accuracy)