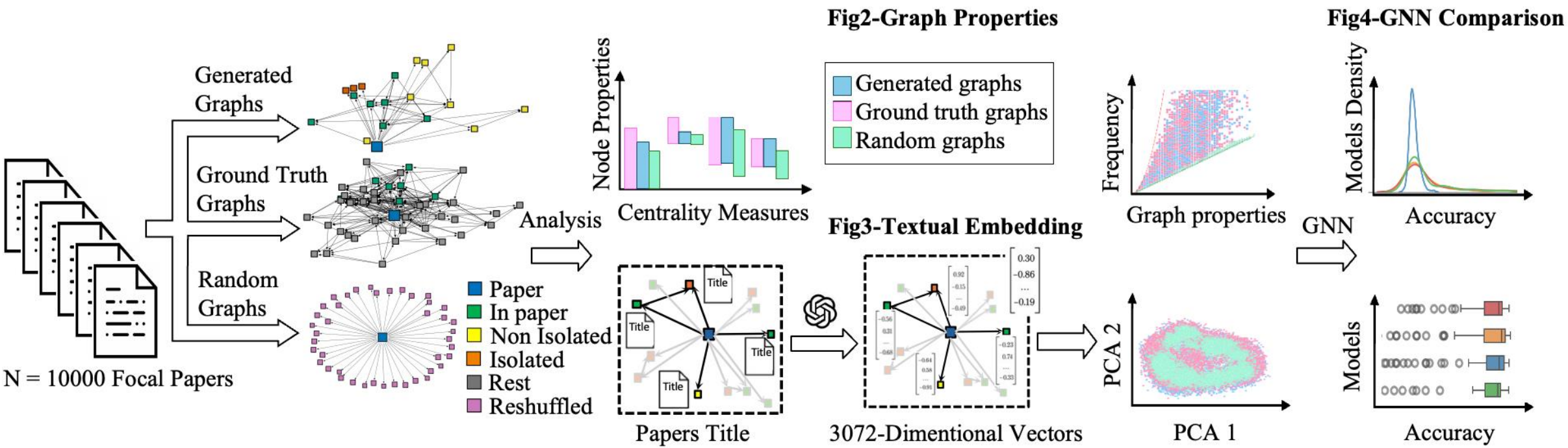
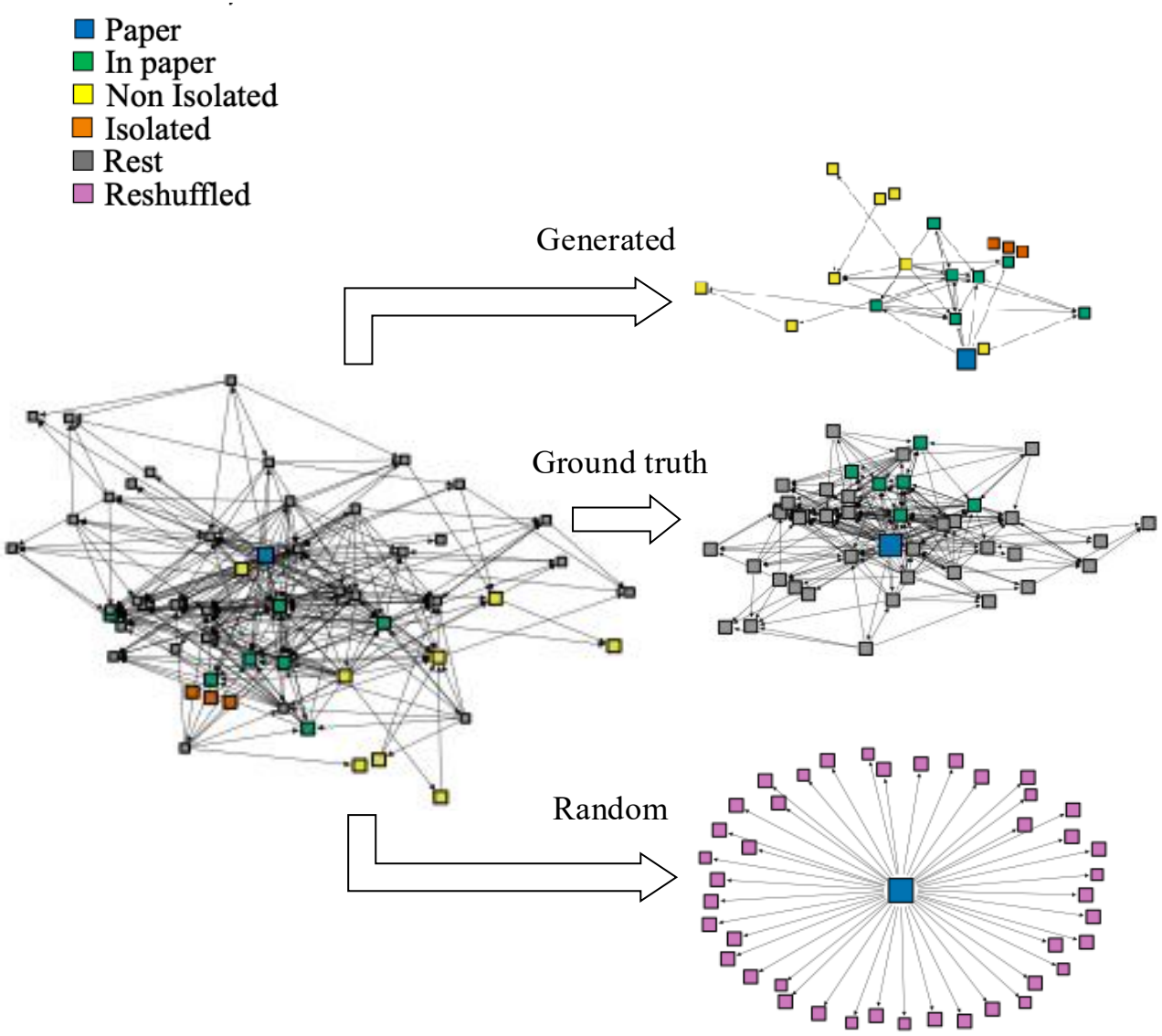
Structurally Human, Semantically Biased:
Detecting LLM-generated references with embeddings and GNNs

Method at a glance



From induced citation graphs → Interpretable topology → Title embeddings → RF and GNNs evaluation.

CONSTRUCTING CITATION NETWORKS



Three graph types (paired per focal paper)

Ground truth

Human reference list

Generated

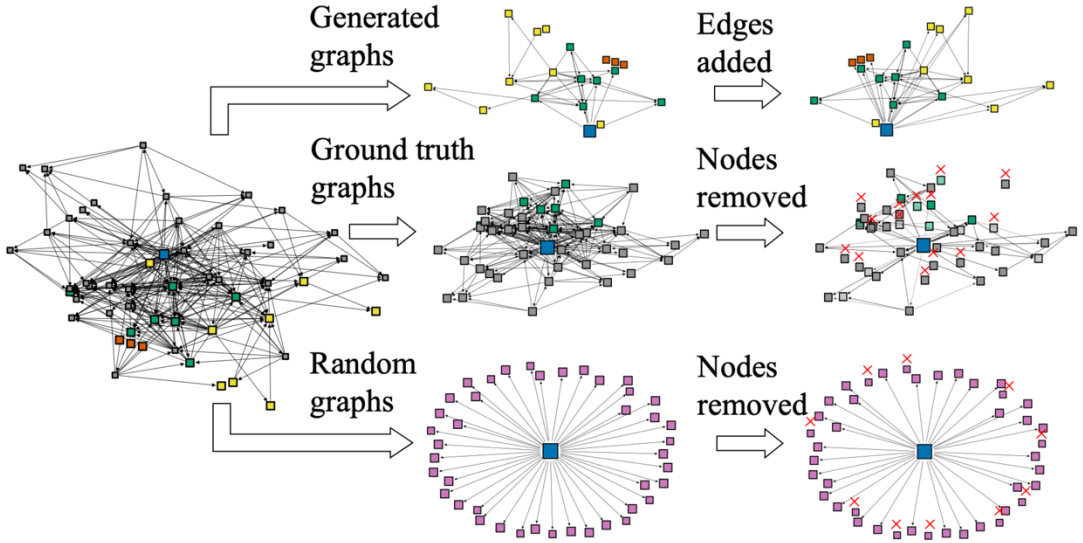
GPT-4o (parametric)

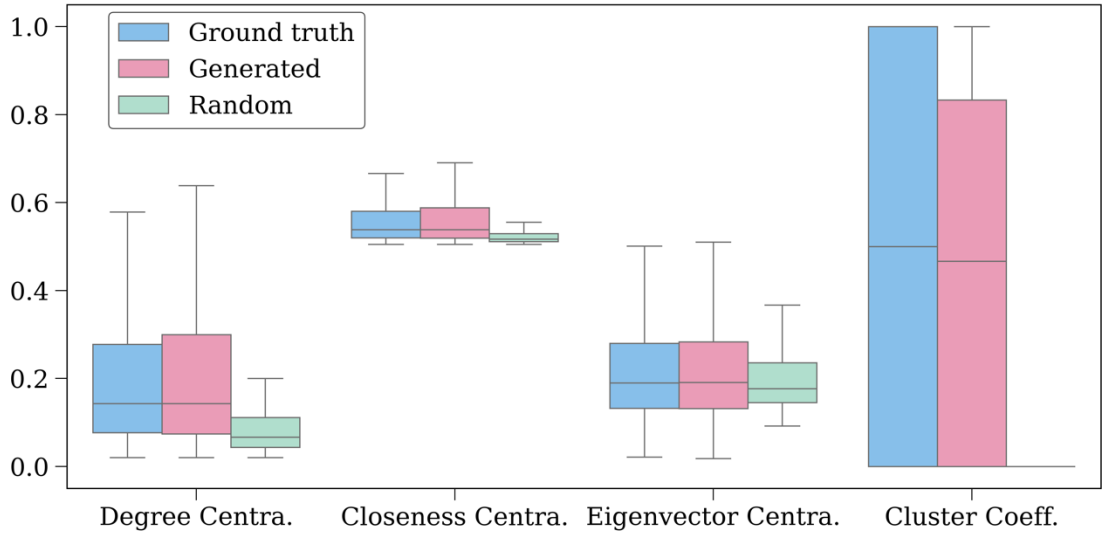
Random baseline

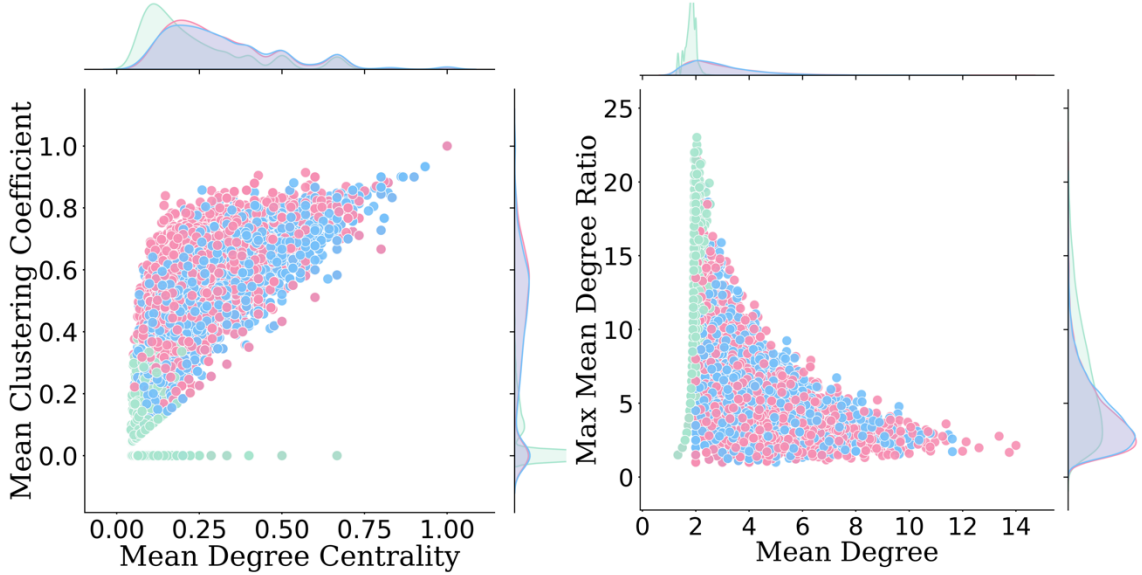
Field-matched reshuffle

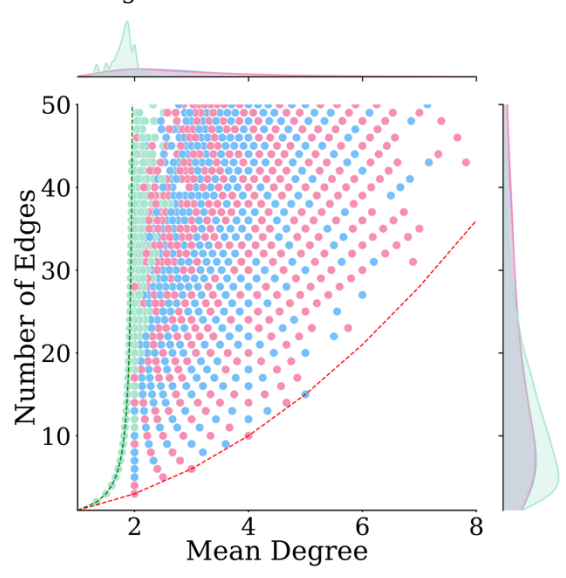
Global Topology

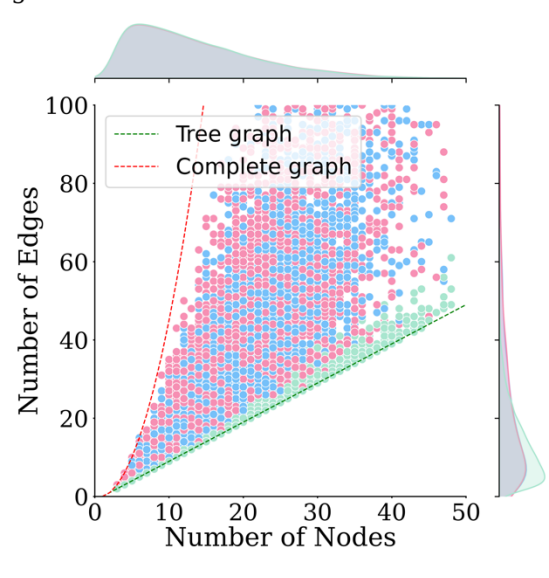
Ground truth and generated graphs overlap almost entirely within structural constraints; random graphs do not.







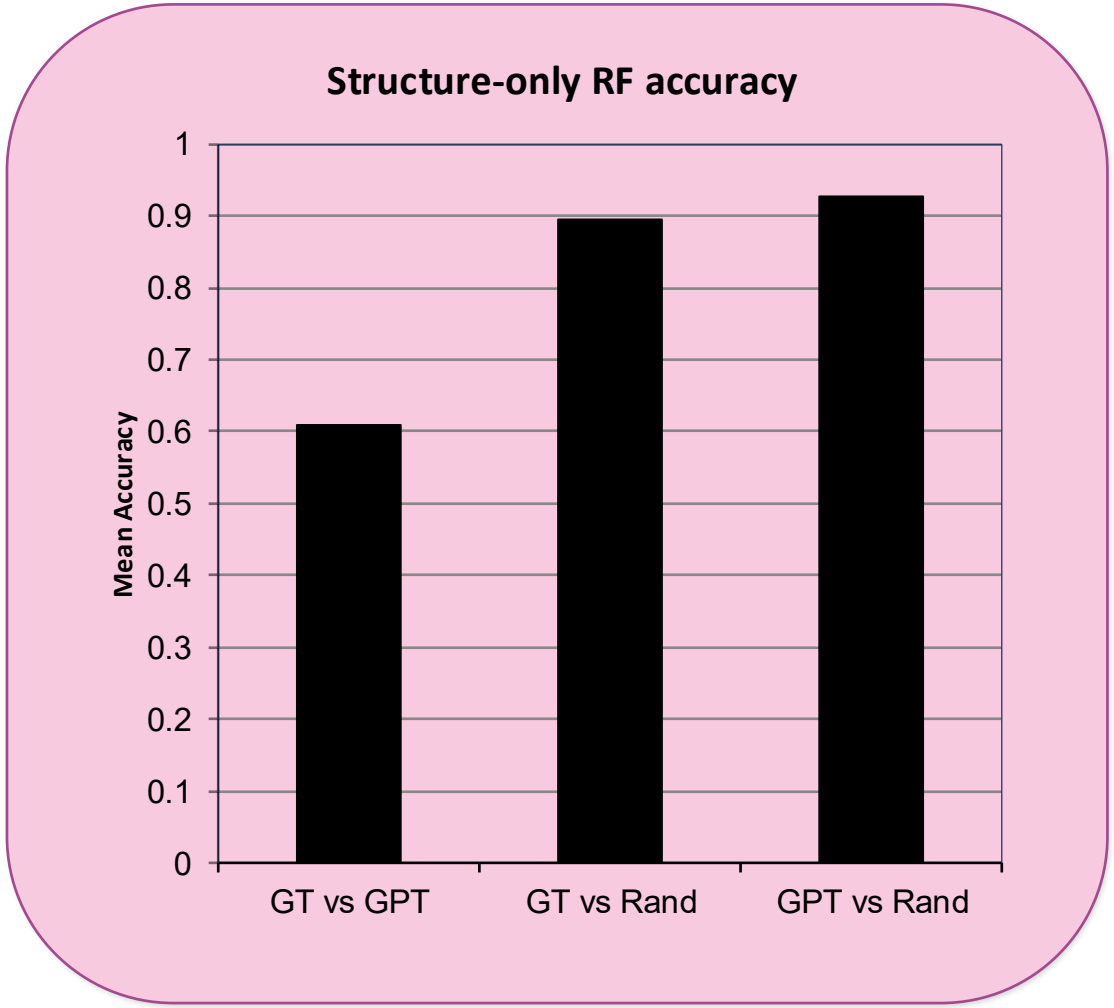




RF on graph properties

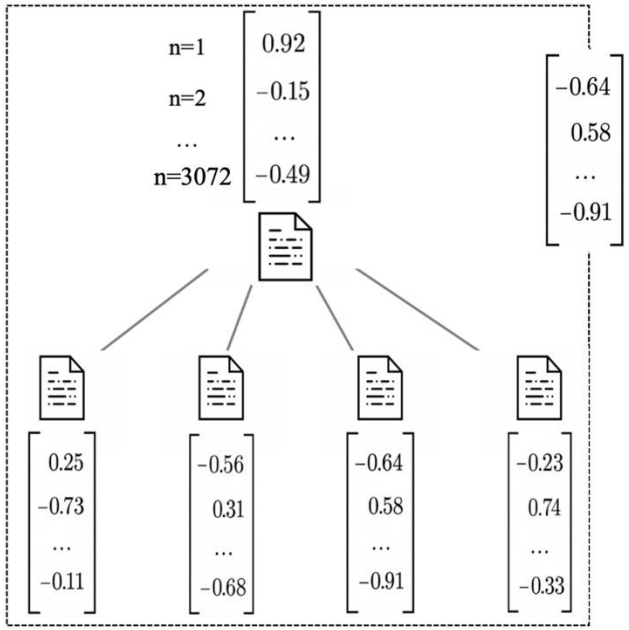
Structure cleanly rejects random baselines, but barely separates GPT vs ground truth

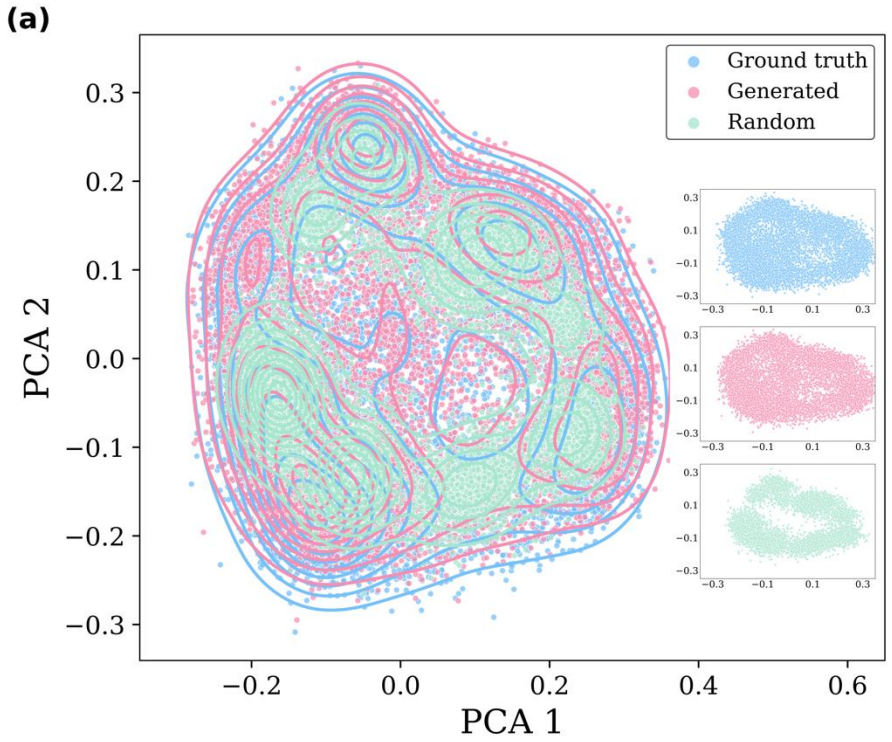
Random Forest on the aggregated graph-level features

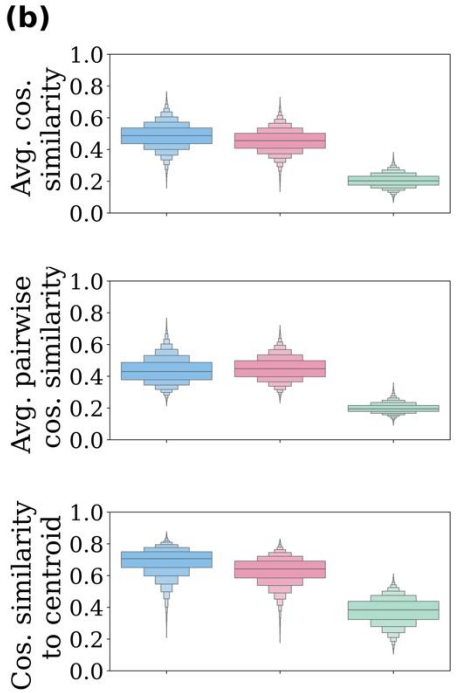


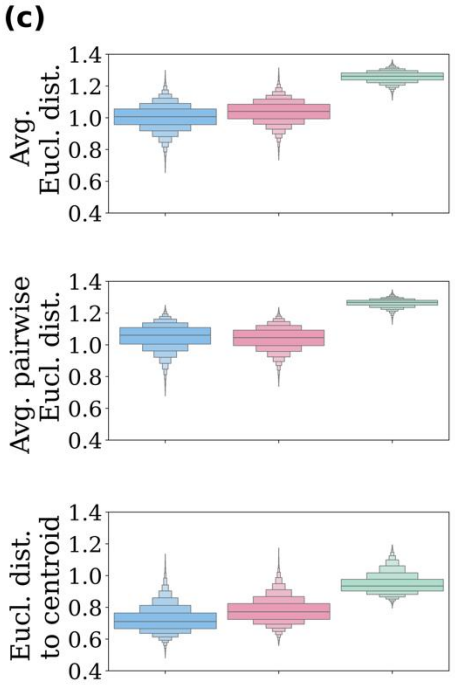
Semantics embeddings

Strong semantic alignment of generated references with ground truth references; Compared to weak alignment in random graphs





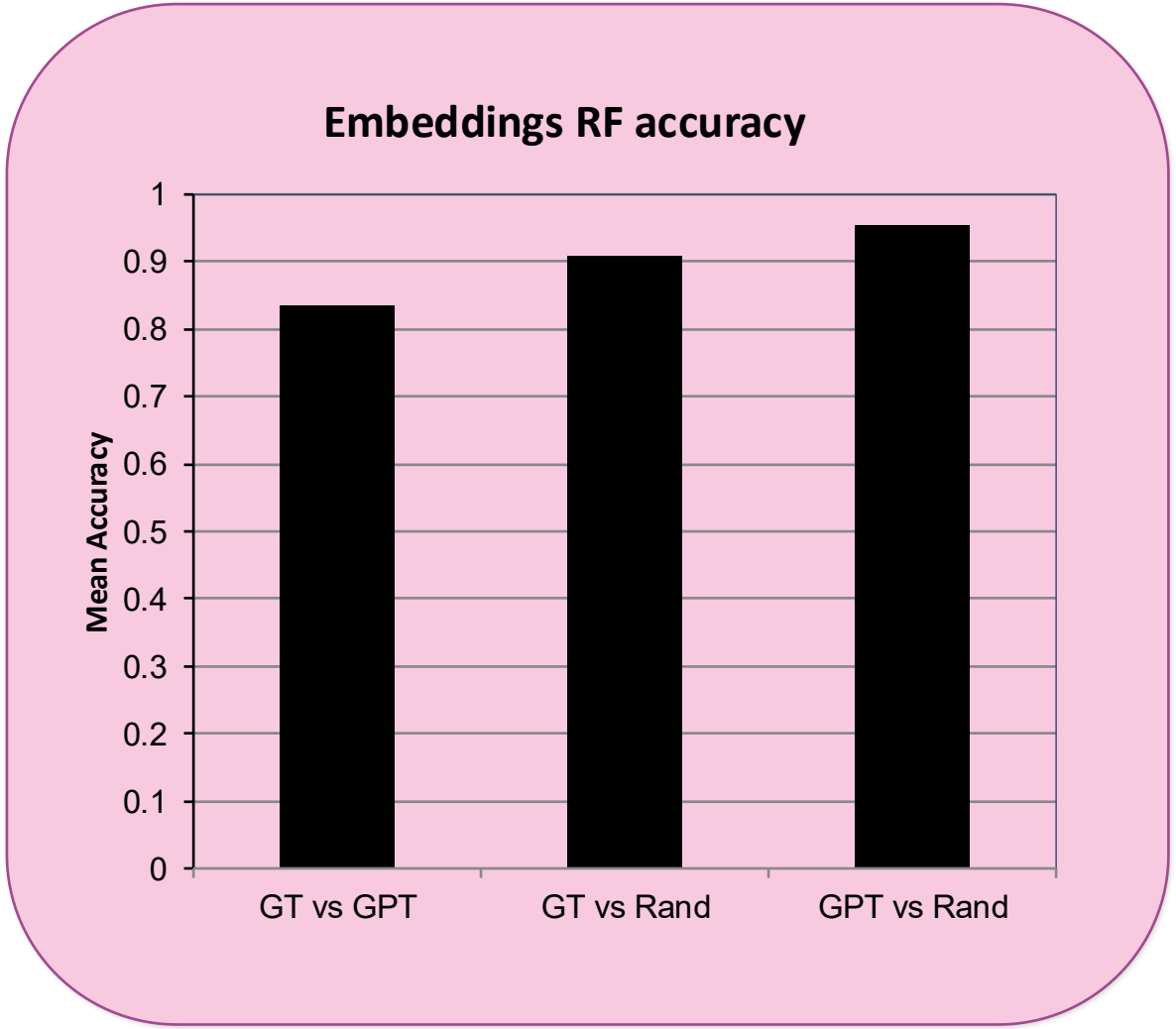




RF on embedding vectors

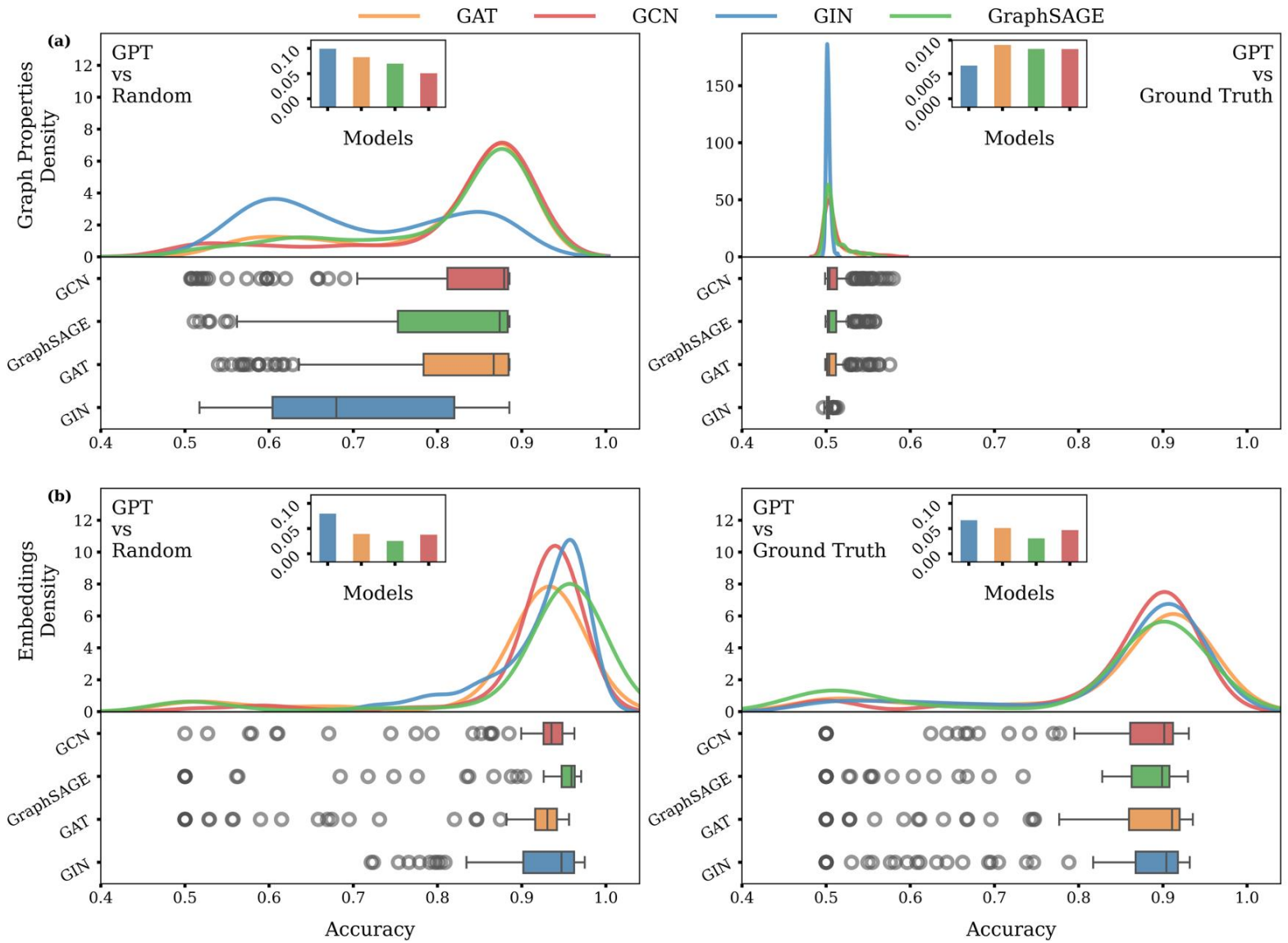
Aggregated title/abstract embeddings increase separability

Random Forest on aggregated embedding vectors



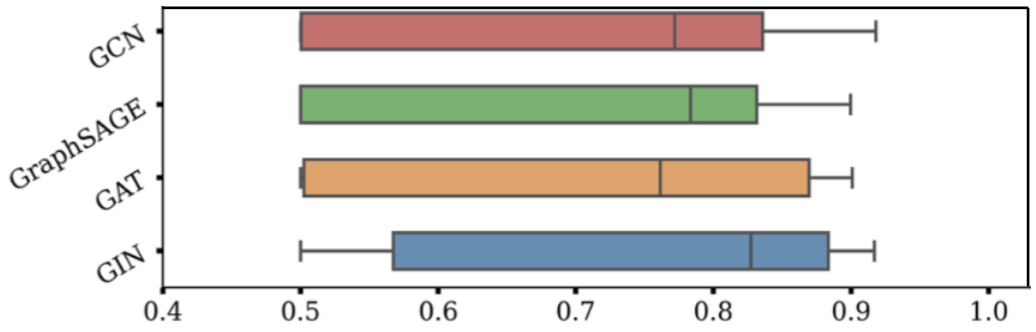
GNN architectures across classification tasks

Distribution of final validation accuracy over the sweep of hyperparameters using different models

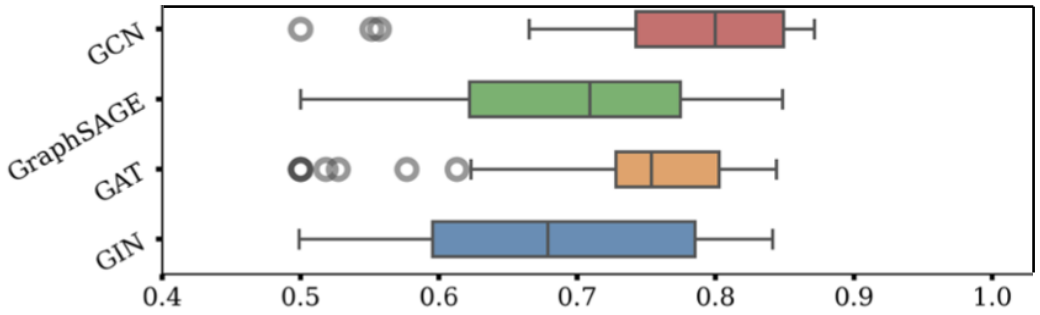


Robustness check

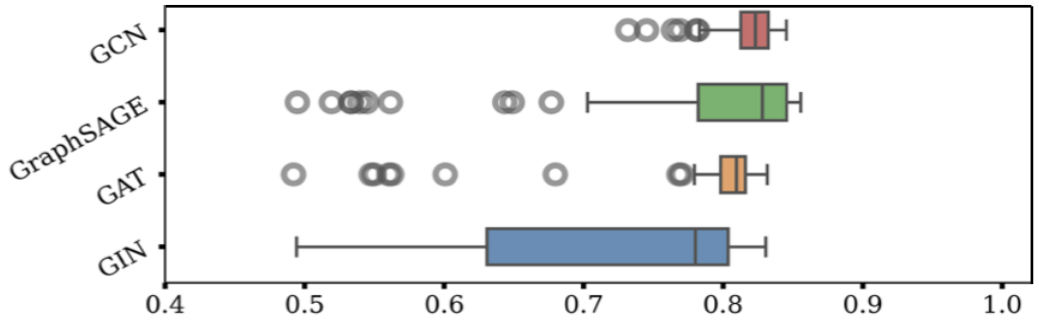
Claude using OpenAI embeddings



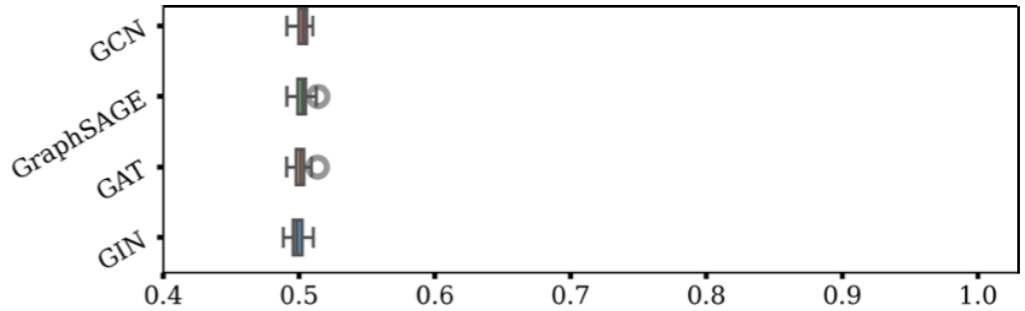
Claude using SPECTER embeddings



GPT using SPECTER embeddings

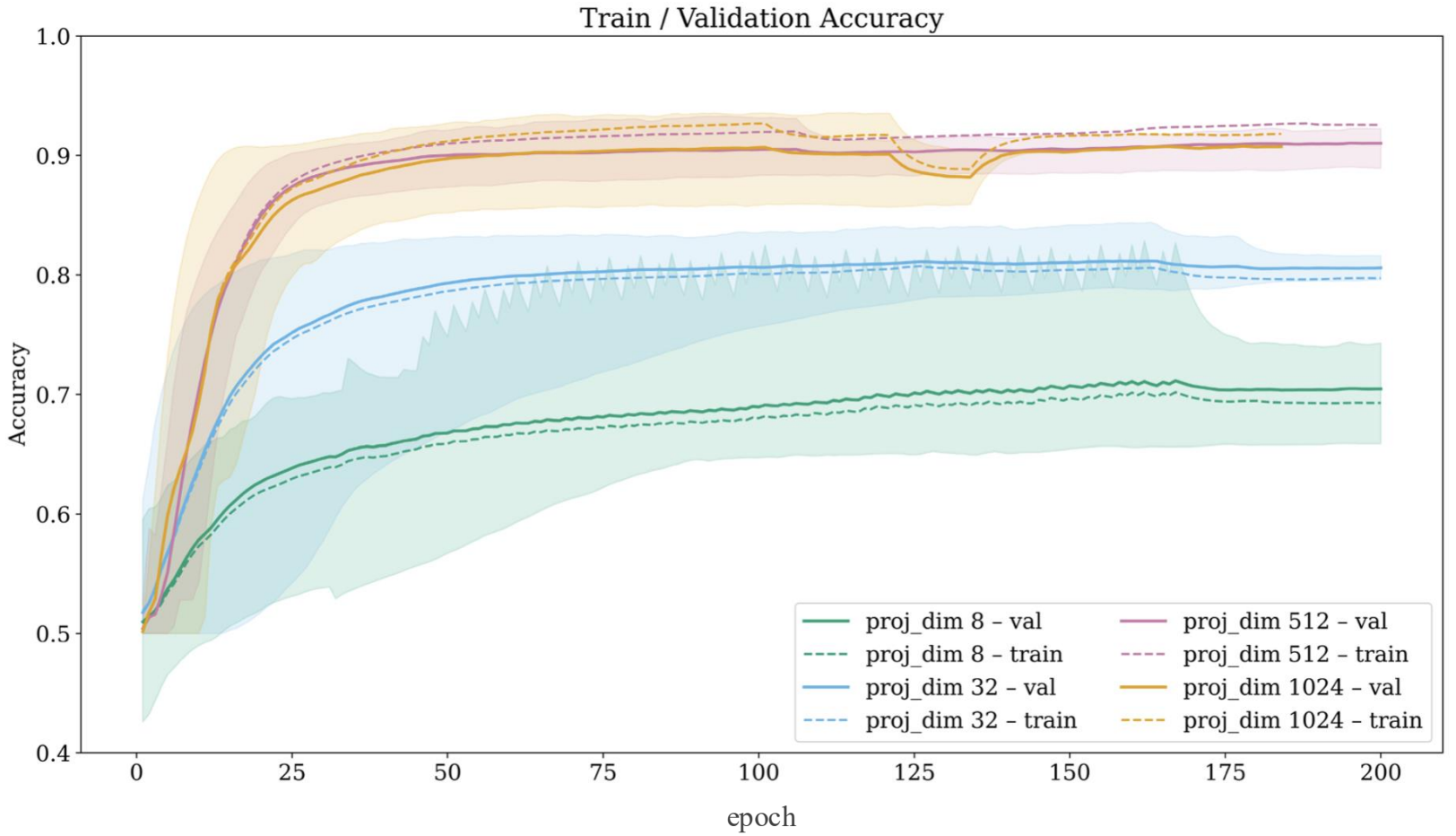


Random embedding



Embedding projection dimension

Dimensionality-controlled ablation



Conclusion and Future Work

Bibliographies can look structurally human, but remain semantically detectable.

Conclusion

- **Structure is not where the signal is**
- **Semantic fingerprints remain detectable**
- Use embedding-based detectors
- The result is **not GPT-4o-specific**

Future Work

- Identify the semantic dimensions driving separability
- Extend beyond titles/abstracts
- Test retrieval-enabled / database-access settings

Melika Mobini
Vrije Universiteit Brussel
Melika.Mobini@vub.be

Vincent Holst
Vrije Universiteit Brussel
Vincent.Thorge.Holst@vub.be

Floriano Tori
Vrije Universiteit Brussel
Floriano.Tori@vub.be

Andres Algaba
Vrije Universiteit Brussel
Andres.Algaba@vub.be

Vincent Ginis
Vrije Universiteit Brussel
SEAS, Harvard University
Vincent.Ginis@vub.be

ICLR

VUB DATA ANALYTICS LABORATORY RESEARCH GROUP

HARVARD
School of Engineering and Applied Sciences