Spectral Manifold Harmonization for Graph Imbalanced Regression

Brenda Nogueira, Meng Jiang, Nitesh V. Chawla, Nuno Moniz

University of Notre Dame

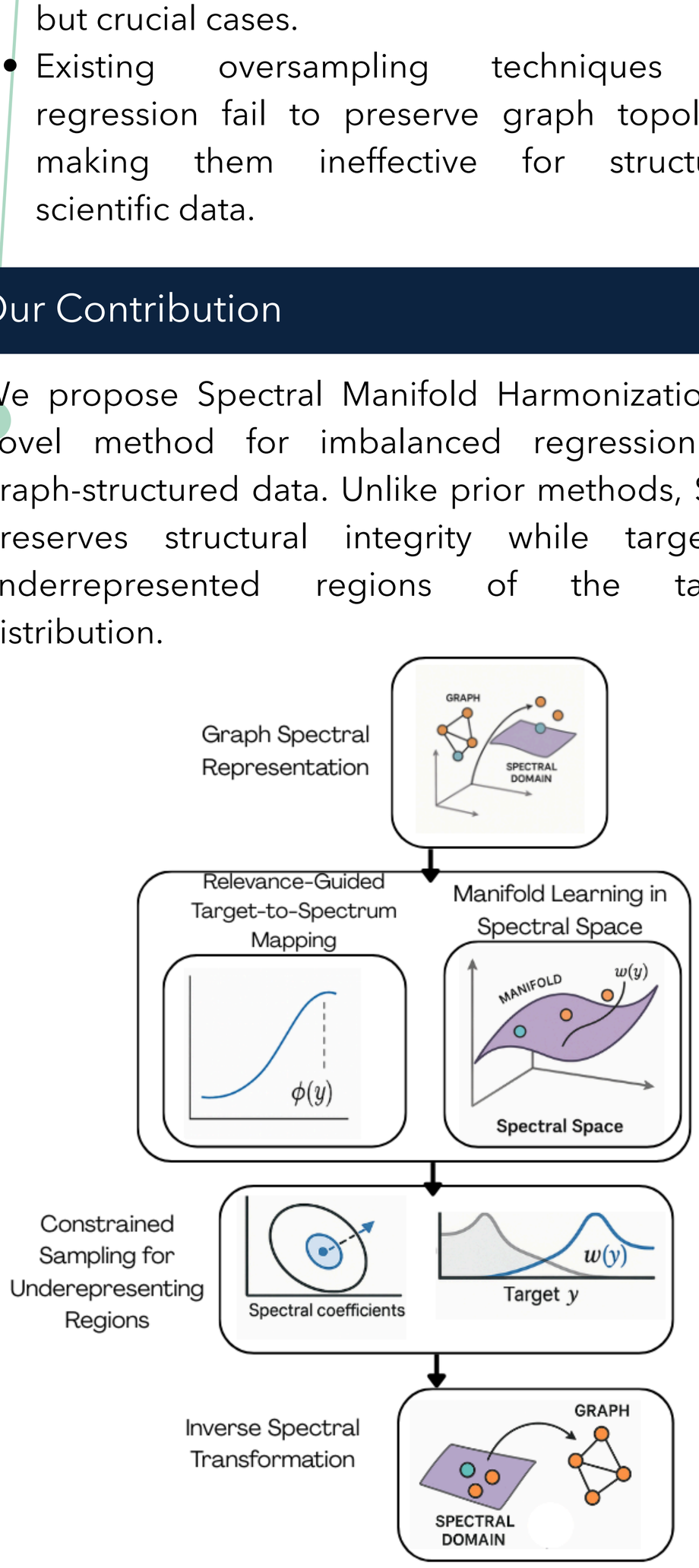
ICML International Conference On Machine Learning

Motivation/Gaps

- Imbalanced regression on graphs is understudied, despite its importance in science (e.g., rare high-potency drugs).
- Most GNN models optimize for average performance, leading to poor accuracy on rare but crucial cases.
- Existing oversampling techniques for regression fail to preserve graph topology, making them ineffective for structured scientific data.

Our Contribution

We propose Spectral Manifold Harmonization, a novel method for imbalanced regression on graph-structured data. Unlike prior methods, SMH preserves structural integrity while targeting underrepresented regions of the target distribution.



Results

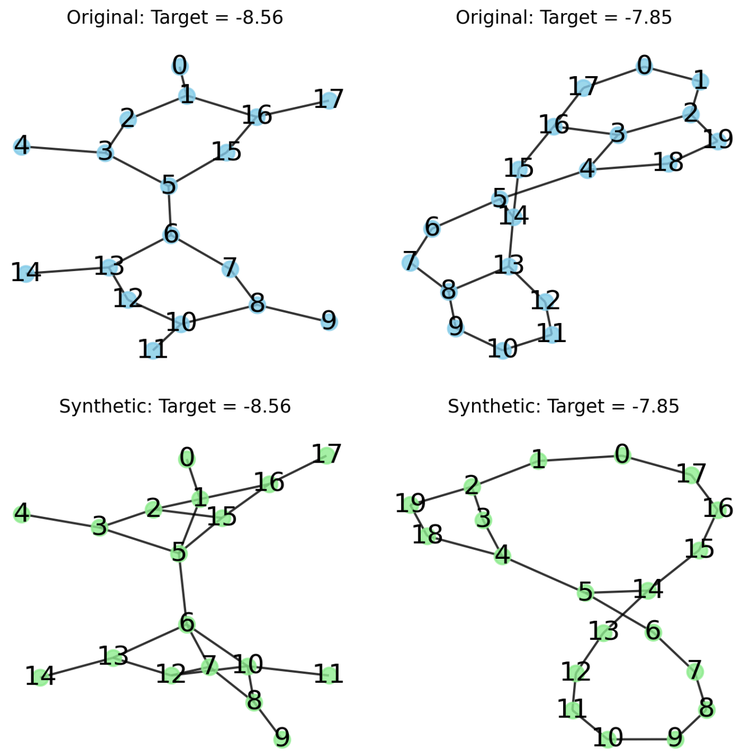


Figure 1. Illustration of graphs selected for augmentation and the corresponding synthetic graphs generated for the ESOL dataset.

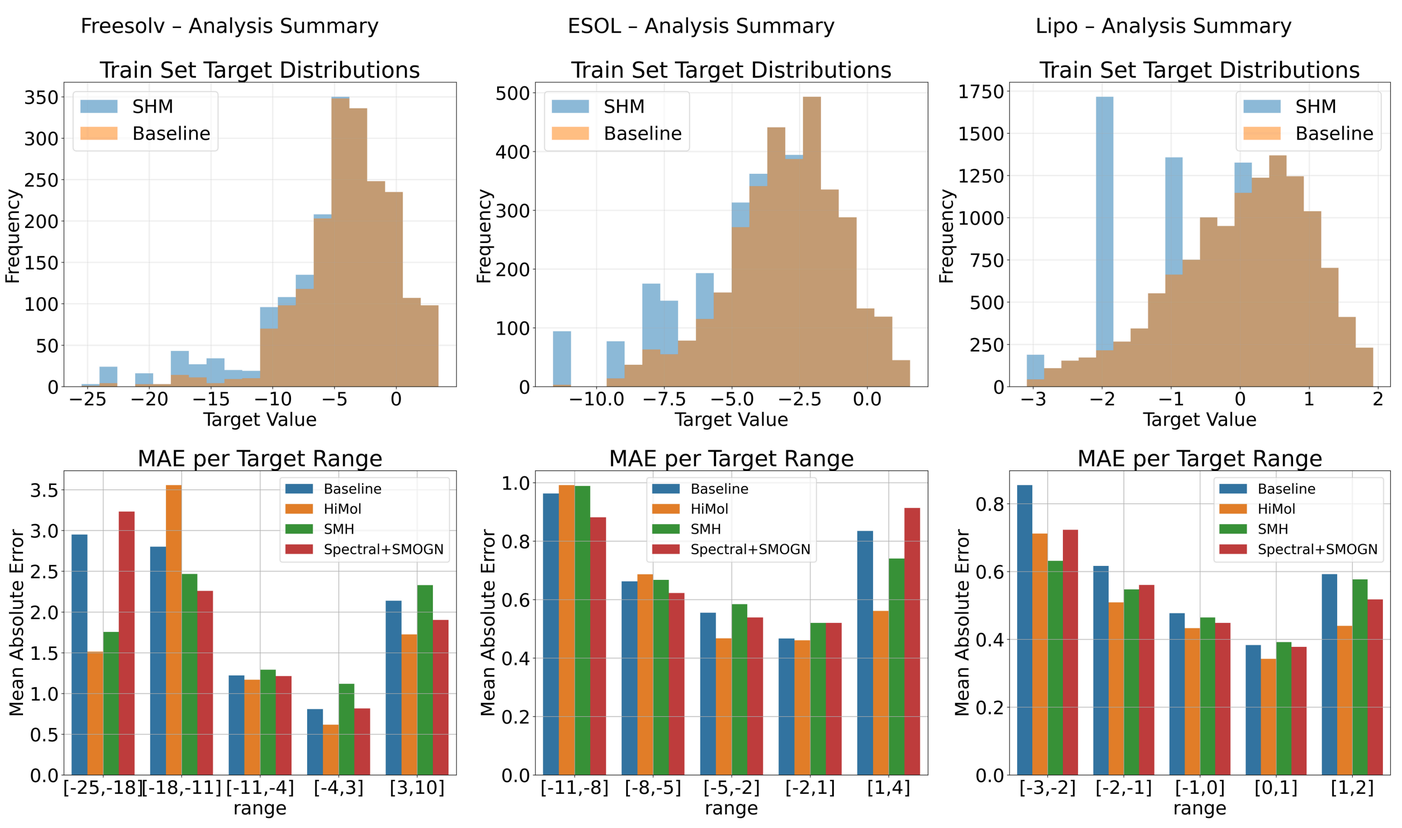


Figure 2. Distribution of the train dataset with and without synthetic augmentation, along with the improvements for each part of the test set domain, for each dataset.

Table1. Experimental results for the FreeSolv, ESOL, and LIPO datasets, using the SERA, MAE, RMSE, and R2 evaluation metrics. Arrows signal the direction for best results, also noted in bold.

Freesolv [1]

Metric	Baseline	SMH	Spectral+SMOGN	HiMol
SERA ↓	0.83 ± 0.90	**0.55 ± 0.35**	0.69 ± 0.58	0.71 ± 0.93
MAE ↓	1.07 ± 0.16	1.25 ± 0.17	1.06 ± 0.14	**0.95 ± 0.17**
RMSE ↓	1.67 ± 0.33	1.81 ± 0.30	1.59 ± 0.32	**1.46 ± 0.41**
R² ↑	0.81 ± 0.07	0.77 ± 0.11	0.83 ± 0.06	**0.85 ± 0.08**

ESOL [1]

Metric	Baseline	SMH	Spectral+SMOGN	HiMol
SERA ↓	0.07 ± 0.03	0.08 ± 0.03	**0.06 ± 0.02**	0.08 ± 0.01
MAE ↓	0.56 ± 0.05	0.59 ± 0.04	0.56 ± 0.02	**0.51 ± 0.02**
RMSE ↓	0.73 ± 0.07	0.77 ± 0.05	0.73 ± 0.04	**0.70 ± 0.02**
R² ↑	0.87 ± 0.03	0.86 ± 0.02	0.88 ± 0.02	**0.89 ± 0.01**

Lipo [1]

Metric	Baseline	SMH	Spectral+SMOGN	HiMol
SERA ↓	0.11 ± 0.03	**0.08 ± 0.01**	0.09 ± 0.02	**0.08 ± 0.01**
MAE ↓	0.49 ± 0.01	0.47 ± 0.02	0.46 ± 0.01	**0.42 ± 0.02**
RMSE ↓	0.66 ± 0.01	0.64 ± 0.02	0.62 ± 0.03	**0.57 ± 0.01**
R² ↑	0.57 ± 0.02	0.60 ± 0.03	0.62 ± 0.04	**0.67 ± 0.01**

Conclusion

- SMH-generated graphs closely match the original in node and edge statistics, with only minor variations in density—confirming structural validity of synthetic samples.
- Significantly improves prediction in underrepresented low-value regions, where training data is scarce, while maintaining stable performance elsewhere, reflected in better SERA scores, compared to Baseline and pretrained model (HiMol).
- SMH achieves more efficient augmentation, especially in critical ranges, when compared to other augmentation strategies (Spectral+SMOGN).

References

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: a benchmark for molecular machine learnng. Chemical science, 9(2):513–530, 2018.

Branco, P., Torgo, L., and Ribeiro, R. P. Smogn: a preprocessing approach for imbalanced regression. In First international workshop on learning with imbalanced domains: Theory and applications

Zang, X., Zhao, X., and Tang, B. Hierarchical molecular graph self-supervised learning for property prediction. Communications Chemistry.

Ribeiro, R.P., Moniz, N. Imbalanced regression and extreme value prediction. Mach Learn 109, 1803–1835 (2020). https://doi.org/10.1007/s10994-020-05900-9