Spectral Manifold Harmonization for Graph Imbalanced Regression

Brenda Nogueira [1] **Meng Jiang** [1] **Nitesh V. Chawla** [1] **Nuno Moniz** [1]

Abstract

Graph-structured data is ubiquitous in scientific domains, where models often face imbalanced learning settings. In imbalanced regression, domain preferences focus on specific target value ranges representing the most scientifically valuable cases; we observe a significant lack of research. In this paper, we present Spectral Manifold Harmonization (SMH), a novel approach for addressing this imbalanced regression challenge on graph-structured data by generating synthetic graph samples that preserve topological properties while focusing on often underrepresented target distribution regions. Conventional methods fail in this context because they either ignore graph topology in case generation or do not target specific domain ranges, resulting in models biased toward average target values. Experimental results demonstrate the potential of SMH on chemistry and drug discovery benchmark datasets, showing consistent improvements in predictive performance for target domain ranges.

1. Introduction

Graph-structured data has become increasingly important in scientific domains, particularly drug discovery, materials science, and genomics. Graph Neural Networks (GNNs) have revolutionized the modeling of such data by operating directly on graph structures, enabling more accurate predictions of molecular properties, material characteristics, and biological interactions, for example. In drug discovery alone, GNNs have demonstrated significant promise for tasks such as property prediction (Xiong et al., 2020), molecular design (Jin et al., 2018), and drug-target interaction prediction (Lim et al., 2019). The pharmaceutical industry has embraced these methods to accelerate the traditionally slow and expensive drug development pipeline, which typically costs over $1 billion and spans more than a decade from discovery to market (Vamathevan et al., 2019).

While considerable research has targeted imbalanced classification problems in graph learning (Almeida et al., 2024; Xia et al., 2024), the regression setting has received comparatively little attention (Ribeiro & Moniz, 2020; Liu et al., 2023). Many crucial scientific problems involve predicting continuous properties where the most valuable cases are rare, e.g., in drug discovery, high-potency compounds represent a tiny fraction of the chemical space but are the most scientifically interesting (Silva et al., 2022). Standard machine learning approaches, including GNNs, typically optimize for average performance across the entire distribution, resulting in poorly performing models on these infrequent but valuable cases. Also, existing oversampling techniques for imbalanced data fail to preserve the complex topological properties inherent in graph-structured scientific data, limiting their effectiveness in these domains.

In this paper, we present Spectral Manifold Harmonization (SMH), a novel approach for tackling imbalanced regression on graph-structured data. SMH (Figure 1) operates in the graph spectral domain—the eigenspace of the graph Laplacian—to generate synthetic graph samples that preserve essential topological properties while focusing on underrepresented target distribution regions. Our approach builds on the concept of relevance in imbalanced regression (Ribeiro & Moniz, 2020), which maps target values to non-uniform domain preferences, assigning higher importance to specific domain ranges. SMH learns a continuous manifold of valid graph structures by establishing a mapping between target regression values and the spectral domain, allowing us to generate new samples with targeted properties.

Novelty. This approach overcomes the limitations of existing oversampling techniques in regression settings by operating in a space that captures the structural properties of graphs, enabling the generation of realistic synthetic examples that address the imbalance problem without distorting the underlying graph topology.

Findings. Experimental results show that SMH considerably improves predictive performance on target ranges in benchmark datasets from drug discovery. Specifically, models trained with SMH-augmented datasets improve the

[1]University of Notre Dame, Notre Dame, Indiana, USA. Correspondence to: Brenda Nogueira <bcruznog@nd.edu>.

Published at Data in Generative Models Workshop: The Bad, the Ugly, and the Greats (DIG-BUGS) at ICML 2025, Vancouver, Canada.

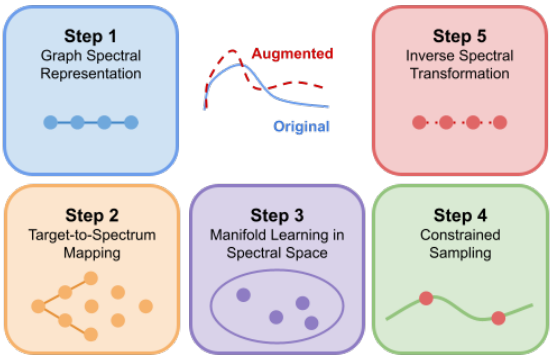

Figure 1: Visual description of the Spectral Manifold Harmonization method's workflow.

accurate prediction of rare compounds, while maintaining or improving performance on average cases. We also demonstrate how synthetic graphs generated by SMH preserve essential structural characteristics of the original data, confirming the effectiveness of our spectral approach.

2. Related Work

The challenge of imbalanced distributions in graph learning tasks has received increasing attention, particularly in scientific domains where rare values are critical. Recent research by Almeida et al. (2024) demonstrated that class imbalances in drug discovery datasets can be effectively addressed through techniques like oversampling and loss function manipulation when using Graph Neural Networks (GNNs). Despite these advances, most approaches operate directly in graph space rather than the spectral domain, limiting their ability to maintain global structural constraints. Bo et al. (2023b) published a comprehensive survey on spectral GNNs, highlighting their unique ability to capture global information and provide better expressiveness than spatial approaches. Wang & Zhang (2022) further analyzed the theoretical expressive power of spectral GNNs, proving that they can produce arbitrary graph signals under specific conditions. However, these methods usually focus on balanced datasets, illustrating the novelty and significance of SMH.

2.1. Spectral Graph Methods

Spectral graph theory has a rich history in machine learning, with applications spanning dimensionality reduction, clustering, and graph signal processing. Recent work in spectral methods includes Specformer (Bo et al., 2023a), combining spectral GNNs with transformer architectures to create learnable set-to-set spectral filters, or the work by (Li et al., 2025) to enhance the scalability of spectral GNNs without decoupling the network architecture, addressing a key limitation in previous approaches. These advanced spectral methods demonstrate improved performance on various graph learning tasks, but do not specifically target the regression setting or leverage the spectral domain for manifold harmonization in imbalanced scenarios. Our SMH method extends these ideas to regression, enabling targeted generation in underrepresented regions while maintaining global graph properties.

2.2. Manifold Learning for Structured Data

Manifold learning principles underpin many approaches to generating synthetic structured data. Recently, Zhong et al. (2024) described how models can be enhanced by incorporating structured knowledge representations and latent manifold embeddings, in the context of knowledge-augmented graph machine learning for drug discovery. Similarly, Baumgartner et al. (2023) demonstrated that incorporating manifold information improves synthetic oversampling techniques for high-dimensional spectral data where standard approaches often fail. Our SMH approach differs from these works by explicitly modeling the regression target-to-spectrum mapping and performing manifold learning in the spectral domain, making it particularly suited for scientific applications with imbalanced regression targets.

2.3. Graph Sampling and Synthesis in Scientific Domains

Due to domain-specific constraints and validity requirements, scientific applications pose unique challenges for graph-based methods. Yao et al. (2024) provided a comprehensive bibliometric analysis of GNN applications in drug discovery, showing significant growth in this area and highlighting the need for methods to handle the inherent data imbalances in these domains. Similarly, Fan et al. (2024) addressed the challenge of overconfident errors in molecular property classification, demonstrating the importance of uncertainty quantification in imbalanced datasets. These approaches focus primarily on classification rather than regression tasks, and do not specifically utilize spectral representations to address imbalance.

On regression tasks, a review on GNNs for predicting synergistic drug combinations (Zhang & Tu, 2023) noted that graph-based models often suffer from imbalanced data distributions, affecting their performance. They emphasized the need for methods to handle such imbalances to improve predictive accuracy effectively. Our SMH method offers a domain-agnostic approach that incorporates scientific validity constraints while focusing on generating underrepresented regions of the target distribution, bridging critical gaps in existing methodologies for imbalanced regression on graph-structured data in scientific applications.

3. Methods: Spectral Manifold Harmonization

Our Spectral Manifold Harmonization (SMH) method addresses imbalanced regression on graph-structured data by learning to generate synthetic graph samples in underrepresented regions of the target distribution while preserving their topological properties. The key insight is that operating in the graph spectral domain allows us to construct a continuous manifold of valid graph structures, making it possible to sample new graphs with targeted properties. SMH integrates the concept of relevance from recent work on imbalanced regression (Ribeiro & Moniz, 2020; Silva et al., 2022) and consists of five main components (Figure 1): we first transform graphs into their spectral representation, learn how target values map to this spectral space with emphasis on relevant regions, model the manifold of valid spectral representations, strategically sample from underrepresented areas, and finally transform back to generate new graph instances that address the imbalance problem.

3.1. Graph Spectral Representation and Relevance Concept

Let $G = (V, E)$ be a graph with $|V| = n$ nodes and a set of edges E. We define the adjacency matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ where $A_{ij} = 1$ if $(i, j) \in E$, 0 otherwise, the degree matrix $\mathbf{D}$ with $D_{ii} = \sum_j A_{ij}$, and the normalized Laplacian $\mathbf{L}_{\text{norm}} = \mathbf{I}-\mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$. The spectral decomposition of $\mathbf{L}_{\text{norm}}$ yields $\mathbf{L}_{\text{norm}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$, where $\mathbf{U} = [u_1, u_2, ..., u_n]$ contains the eigenvectors and $\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$ contains the eigenvalues with $0 = \lambda_1 \leq \lambda_2 \leq ... \leq \lambda_n \leq 2$. For any graph signal $\mathbf{x} \in \mathbb{R}^n$, its Graph Fourier Transform (GFT) is given by $\hat{\mathbf{x}} = \mathbf{U}^T\mathbf{x}$, where $\hat{\mathbf{x}}$ represents the signal in the spectral domain.

A key concept in addressing imbalanced regression is relevance, which maps target values to non-uniform domain preferences (Ribeiro & Moniz, 2020). In this context, a continuous, domain-dependent relevance function $\phi(Y) : \mathcal{Y} \to [0, 1]$ expresses the application-specific bias concerning the target variable $\mathcal{Y}$. A domain expert ideally defines the relevance function for the specific task where the expert inputs information on the available target value-relevance pairs, i.e., which value is considered low or high-relevance. When this information is unavailable, the function can be interpolated from boxplot-based statistics where extreme values are considered high-relevance and the distribution median is considered the lowest point of relevance.

3.2. Relevance-Guided Target-to-Spectrum Mapping

Given a dataset $\mathcal{D} = \{(G_i, y_i)\}_{i=1}^N$ of graph-label pairs, we learn a parameterized function $f_\theta : \mathbb{R} \to \mathbb{R}^k$ that maps regression target values to spectral coefficients, where $k < n$ is the number of significant eigenmodes. The mapping function is implemented as a neural network:

$$f_\theta(y) = \mathbf{W}_L{\cdot}\sigma(\mathbf{W}_{L-1}{\cdot}\sigma(\cdots\sigma(\mathbf{W}_1{\cdot}y{+}\mathbf{b}_1)\cdots){+}\mathbf{b}_{L-1}){+}\mathbf{b}_L \tag{1}$$

where σ is a non-linear activation function, and $\mathbf{W}_l, \mathbf{b}_l$ are learnable parameters. We incorporate the relevance concept into our optimization objective by weighting the loss according to the importance of each target value:

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N} \phi(y_i) \cdot \|\mathbf{s}_i - f_\theta(y_i)\|^2 + \alpha \cdot \Omega(\theta) \tag{2}$$

where $\mathbf{s}_i = \mathbf{U}_i^T\mathbf{x}_i$ are spectral coefficients of graph G_i, α is a regularization parameter, and $\Omega(\theta)$ is a regularization term. This relevance-weighted loss function ensures that the model focuses more on learning the mapping for high-relevance target values.

3.3. Manifold Learning in Spectral Space

We model the distribution of spectral coefficients conditioned on target values as a multivariate Gaussian:

$$p(\mathbf{s}|y) = \mathcal{N}(\mu(y), \Sigma(y)) \tag{3}$$

where $\mu(y) = f_\theta(y)$ and $\Sigma(y)$ is estimated using a relevance-weighted covariance:

$$\Sigma(y) = \sum_{i=1}^{N} w_i(y) \cdot (\mathbf{s}_i - \mu(y))(\mathbf{s}_i - \mu(y))^T \tag{4}$$

with weights determined by target similarity:

$$w_i(y) = \frac{K(y, y_i)}{\sum_{j=1}^{N} K(y, y_j)} \tag{5}$$

where $K(y, y_i) = \exp(-\gamma(y - y_i)^2)$ is a Gaussian kernel. This weighting scheme ensures that the manifold captures the variability in of each region more accurately when modeling the covariance structure.

3.4. Constrained Sampling for Underrepresented Regions

To address target distribution imbalance, we first estimate the density $p(y)$ using kernel density estimation:

$$p(y) = \frac{1}{Nh}\sum_{i=1}^{N} K\Big(\frac{y - y_i}{h}\Big) \tag{6}$$

where K is a kernel function and h is the bandwidth parameter. We define a sampling weight function that combines both the inverse density and the relevance:

$$w(y) = \phi(y) \cdot (p(y) + \epsilon)^{-1} \tag{7}$$

where ϵ is a small constant to prevent division by zero. This function prioritizes regions that are underrepresented (low density) and highly relevant. To generate new samples, we:

1. Sample target values y_{new} with probability proportional to $w(y)$
2. Generate spectral coefficients $\mathbf{s}_{\text{new}} \sim \mathcal{N}(\mu(y_{\text{new}}), \Sigma(y_{\text{new}}))$

3.5. Inverse Spectral Transformation

Finally, to reconstruct graphs, given $\mathbf{s}$, we:

1. Reconstruct spectral representation $\hat{\mathbf{x}} = [\mathbf{s}_{\text{valid}}, \mathbf{0}]$
2. Apply inverse GFT: $\tilde{\mathbf{x}} = \mathbf{U}\hat{\mathbf{x}}$
3. Construct adjacency matrix: $\tilde{A}_{ij} = \sigma(\tilde{\mathbf{x}}_i \cdot \tilde{\mathbf{x}}_j)$, where σ is a sigmoid function

The resulting graph $\tilde{G}$ preserves essential topological properties while targeting underrepresented yet relevant regions of the target distribution, effectively augmenting the training set to improve regression performance on rare but valuable cases. By integrating the concept of relevance throughout our method, we ensure that the synthetic samples generated by SMH focus specifically on the regions of the target space that are most important for the application domain.

4. Experiments

In this section, we evaluate the effectiveness of SMH in generating synthetic samples that preserve key structural patterns from the original molecular dataset and improve prediction performance in domain-relevant target value ranges.

We address the following research questions:

RQ1 Do synthetic graphs generated by SMH follow the molecular structure patterns of the original dataset?

RQ2 Does the use of SHM improve predictive accuracy in target ranges considered scientifically important?

RQ3 How do SMH's that focus on specific domain regions impact the overall performance?

RQ4 How does manifold learning and constrained sampling perform in comparison with traditional augmentation?

RQ5 How does SMH perform in comparison with pre-trained models?

These questions guide our analysis of the structural fidelity of generated samples and the practical impact of SMH on regression performance across diverse benchmarks.

4.1. Methods

We converted the SMILES into a networkx (Hagberg & Conway, 2020) graph to build the spectral manifold harmonization space. Then, we used XGBoost to train a model to predict the eigenvalues from a given target. For property prediction, we then convert the networkx graph format for *PyTorch Geometric* data format and input it in a Graph Isomorphism Network (GIN) (Xu et al., 2018), as it has emerged as a powerful tool for graph-based machine learning tasks due to of its capability to effectively differentiate between different graph structures, using MSE as a loss function. The hyperparameter is presented in Appendix A with a 5-fold cross-validation. We also compared our relevance-guided target-to-spectrum transformation and constrained sampling approach with SMOGN (Branco et al., 2017). For the Spectral+SMOGN baseline, we first compute the spectral representation as described in Section 3.1, and then apply the SMOGN method. The inverse transformation used for decoding remains the same. The code is available at: *https://github.com/brendacnogueira/smh-graph-imbalance.git*. To compare with a pre-trained model, we used HiMol (Zang et al., 2023), which is a framework to learn molecule graph representation for property prediction.

4.2. Data

Our experimental evaluation focuses on molecular data, using regression tasks from MoleculeNet (Wu et al., 2018): ESOL, FreeSolv, and Lipophilicity (Lipo). The datasets are briefly described in Table 1. The datasets exhibit a long-tailed distribution toward the lower end of the property range, and we define our relevance function to assign higher importance to these.

Table 1: Summary of Molecular Property Datasets

Dataset	# of Compounds	Description
ESOL	1,128	Water solubility dataset
FreeSolv	642	Hydration free energy of small molecules in water
Lipophilicity	4,200	Octanol/water distribution coefficient of molecules

5. Results and Discussion

This section addresses the research questions raised in Section 3, specifically concerning SMH's ability to generate synthetic graphs and model performance when using the SMH method for generating and leveraging such data.

5.1. Synthetic Generated Graphs

An illustration of the graphs selected for augmentation and the corresponding synthetic graphs generated using the approach described in Section 3 is presented in Figure 2. Results show that the generated samples follow the molecular structure patterns of the original dataset. Importantly, they are not simple copies but exhibit structural variations, indicating that the method produces diverse, meaningful graphs.

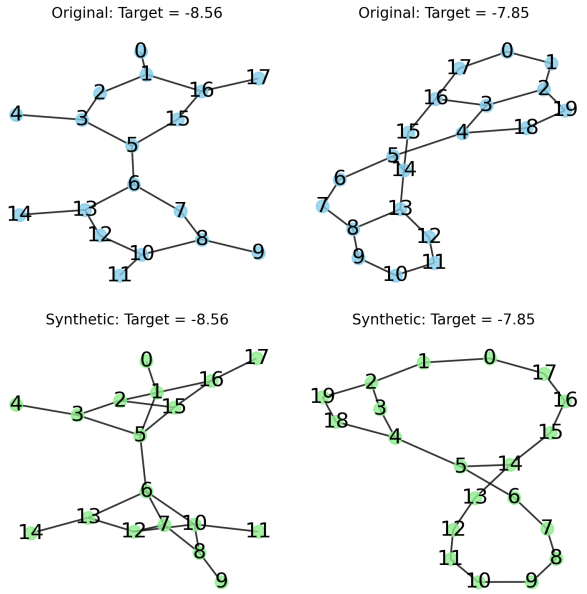

Figure 2: Illustration of graphs selected for augmentation and the corresponding synthetic graphs generated using the approach described in Section 3, for the ESOL dataset.

A comparison of the mean and standard deviation of node and edge counts between the original and synthetic graphs is provided in Table 2. The number of nodes remains nearly identical across SHM and SMOGN. Minor differences are observed in the number of edges and graph density, with the synthetic graphs showing slightly lower mean values. However, these differences remain within an acceptable range, supporting the validity of the generated graphs (**RQ1**). Further validation on the generated graphs can be addressed.

5.2. Model Performance

The experimental results are reported in Table 3, and Figure 3 illustrates each dataset's improvement across different domain regions. The results show noticeable improvements in the lower range of the domain (**RQ2**), where our augmentation is focused and where training data is scarce, with minimal or no degradation in the higher range (**RQ3**). This results in an improvement in the SERA evaluation metric

Table 2: Comparison of Mean and Standard Deviation of Node and Edge Counts Between Original and Synthetic Graphs.

Data set		Node	Edge	Density
FreeSolv	Baseline	8.7 ± 4.19	8.4 ± 4.79	0.3 ± 0.15
	SMH	10.1 ± 3.10	5.0 ± 4.63	0.1 ± 0.12
	Spectral+SMOGN	10.6 ± 2.72	10.8 ± 3.69	0.2 ± 0.07
ESOL	Baseline	13.3 ± 6.93	13.7 ± 8.00	0.2 ± 0.13
	SMH	20.5 ± 4.40	11.3 ± 11.42	0.1 ± 0.06
	SMOGN	20.8 ± 3.43	22.6 ± 3.71	0.1 ± 0.02
Lipo	Baseline	27.1 ± 7.34	29.5 ± 8.12	0.1 ± 0.03
	SMH	23.8 ± 12.69	14.5 ± 14.28	0.1 ± 0.05
	Spectral+SMOGN	23.7 ± 16.59	30.3 ± 23.64	0.1 ± 0.08

and similar results in other metrics. When compared to Spectral+SMOGN, our method improves performance on the most relevant ranges, demonstrating the effectiveness of manifold learning and constrained sampling in generating augmented graphs and their potential for further improvement (**RQ4**). In comparison with a pre-trained model, our approach demonstrates very comparable results with significant improvements in the low range part of the domain (**RQ5**).

Table 3: Experimental results for the FreeSolv, ESOL, and LIPO datasets, using the SERA, MAE, RMSE, and R^2 evaluation metrics. Arrows signal the direction for best results, also noted in bold.

FreeSolv				
Metric	**Baseline**	**SHM**	**Spectral+SMOGN**	**HiMol**
SERA ↓	0.83 ± 0.9	$\mathbf{0.55 \pm 0.35}$	0.69 ± 0.58	0.71 ± 0.93
MAE ↓	1.07 ± 0.16	1.25 ± 0.17	1.06 ± 0.14	$\mathbf{0.95 \pm 0.17}$
RMSE ↓	1.67 ± 0.33	1.81 ± 0.3	1.59 ± 0.32	$\mathbf{1.46 \pm 0.41}$
R^2 ↑	0.81 ± 0.07	0.77 ± 0.11	0.83 ± 0.06	$\mathbf{0.85 \pm 0.08}$
ESOL				
Metric	**Baseline**	**SHM**	**Spectral+SMOGN**	**HiMol**
SERA ↓	0.07 ± 0.03	0.08 ± 0.03	$\mathbf{0.06 \pm 0.02}$	0.08 ± 0.01
MAE ↓	0.56 ± 0.05	0.59 ± 0.04	0.56 ± 0.02	$\mathbf{0.51 \pm 0.02}$
RMSE ↓	0.73 ± 0.07	0.77 ± 0.05	0.73 ± 0.04	$\mathbf{0.7 \pm 0.02}$
R^2 ↑	0.87 ± 0.03	0.86 ± 0.02	0.88 ± 0.02	$\mathbf{0.89 \pm 0.01}$
Lipo				
Metric	**Baseline**	**SHM**	**Spectral+SMOGN**	**HiMol**
SERA ↓	0.11 ± 0.03	$\mathbf{0.08 \pm 0.01}$	0.09 ± 0.02	$\mathbf{0.08 \pm 0.01}$
MAE ↓	0.49 ± 0.01	0.47 ± 0.02	0.46 ± 0.01	$\mathbf{0.42 \pm 0.02}$
RMSE ↓	0.66 ± 0.01	0.64 ± 0.02	0.62 ± 0.03	$\mathbf{0.57 \pm 0.01}$
R^2 ↑	0.57 ± 0.02	0.6 ± 0.03	0.62 ± 0.04	$\mathbf{0.67 \pm 0.01}$

6. Conclusion

In this work, we introduced Spectral Manifold Harmonization (SMH), a novel method for addressing the challenge of imbalanced regression on graph-structured data. By generating synthetic samples in the spectral domain of graphs, SMH maintains topological integrity while focusing learning on underrepresented but domain-relevant target value regions. Our approach bridges a critical gap in the liter-

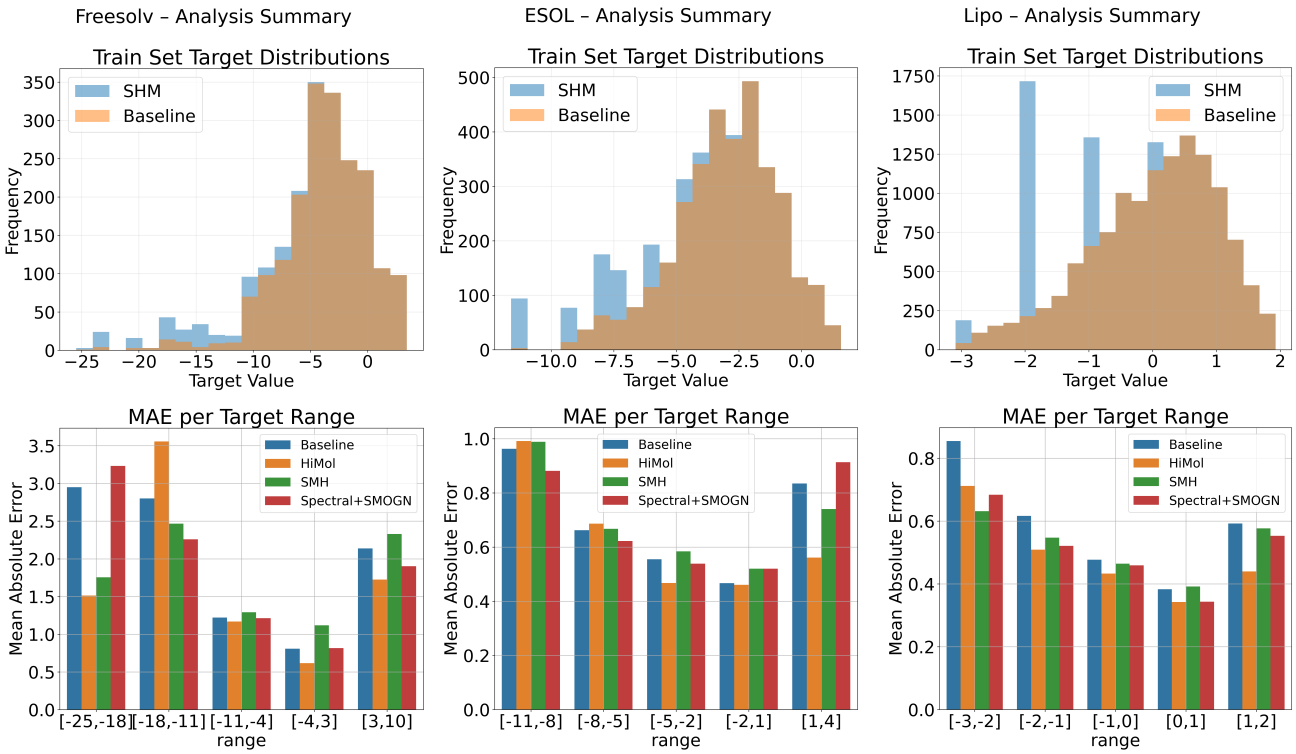

Figure 3: Distribution of train dataset with and without synthetic augmentation, along with the improvements for each part of the test set domain, for each dataset.

ature by combining domain-specific relevance modeling with structure-preserving augmentation, enabling improved predictive performance in settings such as drug discovery where rare cases are of great interest.

Experimental results on benchmark datasets demonstrate that models trained with SMH-augmented data outperform conventional approaches, particularly in low-frequency target regions, without sacrificing performance elsewhere. Structural analyses confirm that generated graphs remain faithful to the original distribution regarding key topological properties. SMH thus offers a principled and effective augmentation strategy for improving learning in scientific domains where data imbalance and structural complexity often limit model effectiveness.

6.1. Future Improvements

Spectral Manifold Harmonization (SMH) has shown strong potential for addressing imbalanced regression on graph-structured data, but several avenues remain for further enhancement. First, integrating domain-specific constraints into the graph generation process could improve the realism and scientific validity of the synthetic graphs. Second, the absence of semantic context in the current synthesis process limits the interpretability and relevance of the generated data, highlighting the need for a hybrid spectral-semantic approach. Future work will also involve evaluating SMH across a wider range of benchmark datasets and predictive models to further optimize performance. Additionally, we plan to conduct more comprehensive comparisons with existing state-of-the-art methods and expand the application domains beyond drug discovery, including areas such as biology and materials science, to better assess the generalizability of our method.

To this end, we aim to develop a hybrid framework that integrates semantic information into the generation and modeling pipeline to further enhance prediction performance and scientific relevance.

References

Almeida, R. L., Maltarollo, V. G., and Coelho, F. G. F. Overcoming class imbalance in drug discovery problems: Graph neural networks and balancing approaches. *Journal of Molecular Graphics and Modelling*, 126:108627, 2024.

Baumgartner, R., Sinn, M., Feurer, F., and Jaeger, S. Manifold-based synthetic oversampling with manifold conformance estimation. *Machine Learning*, 2023.

Bo, D., Shi, C., Wang, L., and Liao, R. Specformer: Spectral

graph neural networks meet transformers. *arXiv preprint arXiv:2303.01028*, 2023a.

Bo, D., Zheng, C., Wang, X., Jiao, P., Zhou, S., Zhang, H., Wei, Z., and Shi, C. A survey on spectral graph neural networks. *arXiv preprint arXiv:2302.05631*, 2023b.

Branco, P., Torgo, L., and Ribeiro, R. P. Smogn: a pre-processing approach for imbalanced regression. In *First international workshop on learning with imbalanced domains: Theory and applications*, pp. 36–50. PMLR, 2017.

Fan, Z., Yu, J., Zhang, X., Chen, Y., Sun, S., Zhang, Y., Chen, M., Xiao, F., Wu, W., Li, X.-N., et al. Reducing overconfident errors in molecular property classification using posterior network. *Patterns*, 2024.

Hagberg, A. and Conway, D. Networkx: Network analysis with python. *URL: https://networkx. github. io*, pp. 1–48, 2020.

Jin, W., Barzilay, R., and Jaakkola, T. Junction tree variational autoencoder for molecular graph generation. *International Conference on Machine Learning*, pp. 2323–2332, 2018.

Li, T., Yin, H., Shi, C., and Lin, W. Large-scale spectral graph neural networks via laplacian sparsification: Technical report. *arXiv preprint arXiv:2501.04570*, 2025.

Lim, J., Ryu, S., Park, K., Choe, Y. J., Ham, J., and Kim, W. Y. Predicting drug-target interaction using a novel graph neural network with 3d structure-embedded graph representation. *Journal of Chemical Information and Modeling*, 59(9):3981–3988, 2019.

Liu, G., Zhao, T., Inae, E., Luo, T., and Jiang, M. Semi-supervised graph imbalanced regression. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, pp. 1453–1465, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599497. URL https://doi.org/10.1145/3580305.3599497.

Ribeiro, R. P. and Moniz, N. Imbalanced regression and extreme value prediction. *Machine Learning*, 109(9): 1803–1835, 2020.

Silva, A., Ribeiro, R. P., and Moniz, N. Model optimization in imbalanced regression. *Lecture Notes in Computer Science*, 13601:1–16, 2022.

Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6):463–477, 2019.

Wang, X. and Zhang, M. How powerful are spectral graph neural networks. *arXiv preprint arXiv:2205.11172*, 2022.

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 learning. *Chemical science*, 9(2):513–530, 2018.

Xia, R., Zhang, C., and Zhang, Y. A novel graph oversampling framework for node classification in class-imbalanced graphs. *Science China Information Sciences*, 67(1):162101, 2024.

Xiong, Z., Wang, D., Liu, X., Zhong, F., Wan, X., Li, X., Li, Z., Luo, X., Chen, K., Jiang, H., et al. Attentive fp: Augmenting graph neural networks with attentive message passing for molecular property prediction. *Journal of Chemical Information and Modeling*, 60(6):2213–2228, 2020.

Xu, K., Hu, W., Leskovec, J., and Jegelka, S. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

Yao, R., Shen, Z., Xu, X., Ling, G., Xiang, R., Song, T., Zhai, F., and Zhai, Y. Knowledge mapping of graph neural networks for drug discovery: a bibliometric and visualized analysis. *Frontiers in Pharmacology*, 15, 2024.

Zang, X., Zhao, X., and Tang, B. Hierarchical molecular graph self-supervised learning for property prediction. *Communications Chemistry*, 6(1):34, 2023.

Zhang, B. and Tu, M. A review on graph neural networks for predicting synergistic drug combinations. *Artificial Intelligence Review*, 2023.

Zhong, Z., Barkova, A., and Mottin, D. Knowledge-augmented graph machine learning for drug discovery: From precision to interpretability. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2024.

A. Model Details.

This section provides additional information about the models used in our experiments. The hyperparameters tested for XGBoost, the property prediction model, and the augmentation strategies are summarized in Table 4. Given the relatively small dimensionality of the eigenvalue vectors (fewer than 50 features), XGBoost outperformed neural networks in our evaluations. Model selection was based on performance on the validation split, using the SERA metric.

Table 4: Hyperparameter search space

Parameter	**Values Tested**	
XGBoost	Number of estimators	10, 50, 100, 250
	Learning rate	0.001, 0.01, 0.1
	Max depth	3, 5, 10
GIN Model	Learning rate	0.01, 0.005, 0.001
	Batch size	16
	Hidden dimension	32, 64
	Number of layers	2, 5
	Epochs	500
SMH	γ	1.0, 0.5
	Augmentation sampling	0.20, 0.15, 0.10
	Binarization cut-off	0.3, 0.2, 0.1
SMOGN	Relevance threshold	0.95, 0.99

We defined a relevance function $\phi(y)$ using the `extremes` method with three control points:

$$\phi(y) = \begin{cases} 1 & \text{if } y = \min(\mathcal{Y}) \\ 0.025 & \text{if } y = \mu = \text{mean}(\mathcal{Y}) \\ 0 & \text{if } y = \max(\mathcal{Y}) \end{cases}$$

where $\mathcal{Y}$ denotes the set of target values in the training data. The relevance function smoothly interpolates between these points to emphasize extreme values.

The training and validation losses are presented in Figure 4.

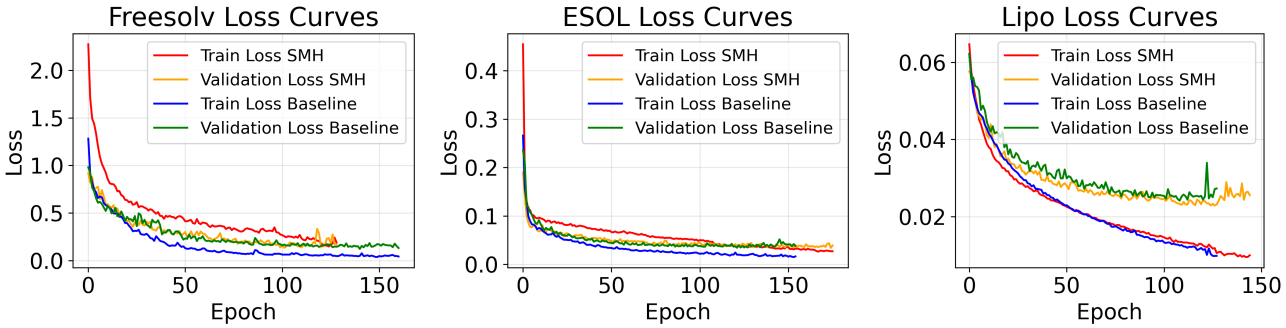

Figure 4: Training and validation performance of property value prediction for each dataset for original and augmented training sets.