

# SPECTRA: SPECTRAL TARGET-AWARE GRAPH AUGMENTATION FOR IMBALANCED MOLECULAR PROPERTY REGRESSION

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

Department of Computer Science  
University of Notre Dame  
Notre Dame, IN, USA

{bcruznog, mjiang2, nchawla, nunomoniz}@nd.edu

## ABSTRACT

In molecular property prediction, the most valuable compounds (e.g., high potency) often occupy sparse regions of the target space. Standard Graph Neural Networks (GNNs) commonly optimize for the average error, underperforming on these uncommon but critical cases, with existing oversampling methods often distorting molecular topology. In this paper, we introduce SPECTRA, a Spectral Target-Aware graph augmentation framework that generates realistic molecular graphs in the spectral domain. SPECTRA (i) reconstructs multi-attribute molecular graphs from SMILES; (ii) aligns molecule pairs via (Fused) Gromov–Wasserstein couplings to obtain node correspondences; (iii) interpolates Laplacian eigenvalues/eigenvectors and node features in a stable shared basis; and (iv) reconstructs edges to synthesize physically plausible intermediates with interpolated targets. A rarity-aware budgeting scheme, derived from a kernel density estimation of labels, concentrates augmentation where data are scarce. Coupled with a spectral GNN using edge-aware Chebyshev convolutions, SPECTRA densifies underrepresented regions without degrading global accuracy. On benchmarks, SPECTRA consistently improves error in relevant target ranges while maintaining competitive overall MAE, and yields interpretable synthetic molecules whose structure reflects the underlying spectral geometry. Our results demonstrate that spectral, geometry-aware augmentation is an effective and efficient strategy for imbalanced molecular property regression.

## 1 INTRODUCTION

Graph-structured data plays a central role in many scientific domains, including drug discovery, materials science, and genomics. These fields produce large volumes of complex, structured information that can be naturally represented as graphs, where nodes correspond to entities (e.g., atoms, molecules, genes) and edges capture their relationships (e.g., chemical bonds, interactions). Graph Neural Networks (GNNs) have transformed the modeling of such data by operating directly on graph structures, enabling state-of-the-art predictions of molecular properties, material characteristics, and biological interactions. In drug discovery, for example, GNNs have been applied to property prediction (Xiong et al., 2020), molecular design (Jin et al., 2018), and drug–target interaction prediction (Lim et al., 2019), with increasing adoption by the pharmaceutical industry to accelerate a development pipeline that typically exceeds \$1 billion and a decade of effort (Vamathevan et al., 2019). Similar advances have been reported in materials science, where GNNs help identify compounds with desirable structural and functional properties (Karamad et al., 2020).

Despite this progress, a fundamental challenge remains largely unsolved: *imbalanced regression* on graphs. While imbalanced classification has received significant attention in graph learning, the regression setting has been comparatively neglected (Almeida et al., 2024; Xia et al., 2024; Ribeiro & Moniz, 2020a; Liu et al., 2023b). Yet, many scientific problems involve continuous targets where the most valuable outcomes are rare. Standard GNNs and other machine learning methods typically optimize for average performance across the full label distribution, which leads to poor accuracy

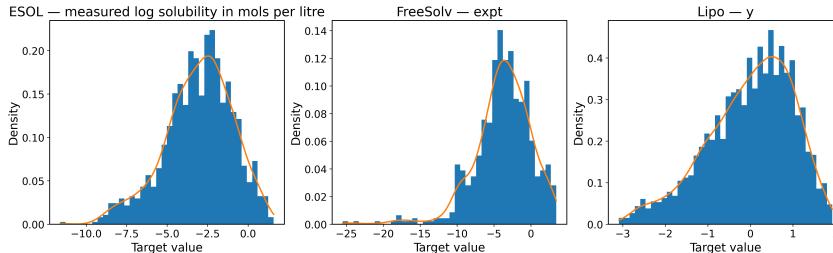


Figure 1: Distribution of target property values across three molecular datasets (ESOL, FreeSolv, and Lipo). Each subplot shows a normalized histogram of the experimental values with a Gaussian kernel density estimate (KDE) overlaid using Scott’s rule-of-thumb bandwidth. These plots highlight the skewness and spread of target distributions, which can influence model training and performance.

in these rare but scientifically important regions. Classical oversampling techniques can mitigate imbalance but often fail to preserve the intricate topological and chemical properties of molecular graphs, limiting their practical effectiveness.

Another limitation lies in *embedding-based augmentation methods*, which often generate synthetic molecules in latent spaces that lack interpretability and offer no guarantees of structural validity. As a result, the augmented samples may not correspond to chemically realistic molecules, hindering trust and practical adoption.

To address these challenges, we propose **SPECTRA**—*Spectral Target-Aware Graph Augmentation for Imbalanced Molecular Property Regression*. SPECTRA introduces a novel approach to oversampling that operates directly in the spectral domain of graphs. Specifically, it leverages the eigenspace of the graph Laplacian to interpolate both Laplacian spectra and node features of matched graphs in a shared spectral basis. This process produces synthetic molecular graphs that are structurally coherent, chemically plausible, and explicitly tailored to underrepresented regions of the target distribution 1. Unlike black-box embedding methods, SPECTRA provides interpretability by generating realistic molecules whose structures can be directly examined, while achieving significantly lower computational cost compared to existing state-of-the-art techniques.

Our contributions can be summarized as follows:

- **Novel methodology.** We introduce a spectral augmentation framework that augments samples in low-density regions of the label space while preserving topological fidelity, overcoming the limitations of existing oversampling techniques in regression.
- **Improved predictive performance.** Across benchmark molecular property datasets, SPECTRA achieves low error on rare compounds without degrading performance on common cases.
- **Interpretability and efficiency.** The synthetic graphs generated by SPECTRA are realistic and chemically meaningful, enabling direct inspection of augmented molecules while maintaining a lower computational footprint compared to competing approaches.

Together, these findings demonstrate that spectral graph augmentation is an effective and interpretable strategy for tackling imbalanced regression in molecular property prediction. The code and dataset are available in <https://anonymous.4open.science/r/SPECTRA-0D3C>

## 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) demonstrates that imbalanced learning in drug discovery datasets can be tackled with techniques such as 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 focus on balanced and classification datasets, illustrating the novelty and significance of SPECTRA.

## 2.1 IMBALANCED LEARNING

Class imbalance has traditionally been addressed through resampling strategies, such as under-sampling majority classes or over-sampling minority classes. SMOTE (Chawla et al., 2002), for instance, generates synthetic minority samples by interpolating labeled data. Alternative approaches include cost-sensitive learning (Cui et al., 2019; Lin et al., 2017), which increases the loss weight of minority classes, and posterior re-calibration (Cao et al., 2019; Menon et al., 2020; Tian et al., 2020), which encourages larger margins for minority predictions.

Imbalanced regression introduces additional challenges because the labels are continuous rather than categorical (Ribeiro & Moniz, 2020a). Several methods from classification have been adapted to this setting. For example, SMOGN Branco et al. (2017) extends SMOTE to regression, while BMSE Ren et al. (2022) adapts logit re-calibration for numerical targets. LDS Yang et al. (2021) smooths the label distribution using kernel density estimation, and RankSim Gong et al. (2022) regularizes the latent space by aligning distances in label and feature space. Other approaches include SERA Ribeiro & Moniz (2020b), which proposes a relevance-aware evaluation metric; SGIR Liu et al. (2023a), which leverages unlabeled graphs to enrich underrepresented label ranges; and SIRN Zong et al. (2024), which combines deviation modeling with adaptive pseudo-label selection. While these methods improve performance in underrepresented regions, they often reduce accuracy in well-represented areas, especially under limited supervision or when relying heavily on pseudo-labeling.

## 2.2 SPECTRAL GRAPH METHODS

Spectral graph theory has 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. Yang et al. (2024) present a spectral-aware augmentation method that selectively perturbs eigenpairs to preserve task-relevant frequency bands in graph contrastive learning. 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 learning in imbalanced scenarios.

## 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. 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.

## 2.4 MOLECULAR GENERATION

Molecular generation has become a central task in drug discovery, aiming to explore chemical space efficiently while ensuring chemical validity and optimizing for desired properties. Early approaches combined variational autoencoders (VAEs), recurrent neural networks (RNNs), and adversarial models to generate novel chemical structures from latent spaces, as in LatentGAN (Prykhodko et al.,

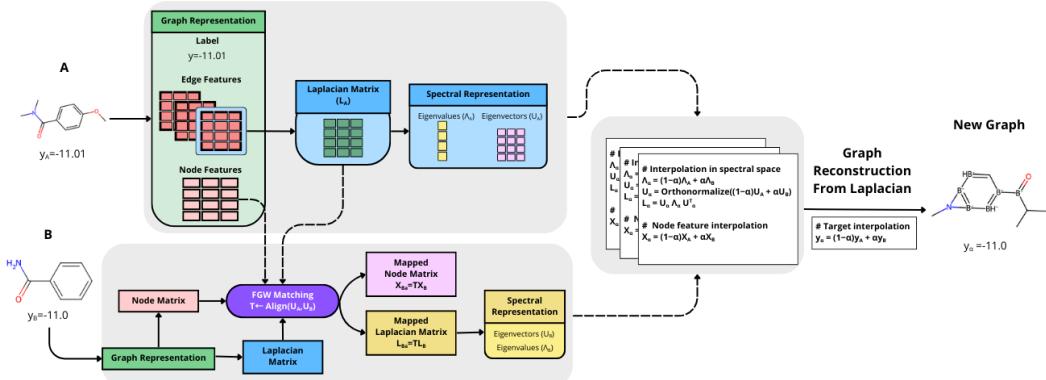


Figure 2: Pipeline of spectral molecular interpolation. Molecular graphs are first aligned via Gromov–Wasserstein matching, after which their three edge-specific Laplacians are decomposed and interpolated in the spectral domain, while node features are projected into the aligned eigenbasis and combined in the same way. Target values are interpolated alongside these representations, producing coherent intermediate graphs that preserve topology while smoothly blending molecular properties and labels to enrich underrepresented regions of the distribution.

2019), which integrated autoencoding with generative adversarial training for de novo molecular design. More recent methods leverage reinforcement learning to incorporate chemical constraints and multi-objective optimization. For example, DeepGraphMolGen (Khemchandani et al., 2020) employs Graph Convolutional Policy Networks to generate molecules while simultaneously optimizing for drug-likeness and synthetic accessibility, whereas MORLD (Jeon & Kim, 2020) integrates reinforcement learning with docking simulations to propose inhibitors directly guided by protein structures. Conditional generative frameworks, such as MGCVAE (Lee & Min, 2022), enable property-conditioned molecular graph generation, allowing for inverse design tasks like optimizing logP or molar refractivity. Beyond purely graph-based approaches, protein-informed generation methods such as DeepTarget (Chen et al., 2023) directly construct candidate molecules from amino acid sequences of target proteins, bridging structural biology with generative chemistry. Despite these advances, most existing models focus on validity, novelty, and property optimization, without explicitly addressing the imbalance of molecular property distributions.

## 2.5 SPECTRA NOVELTY

Our SPECTRA method introduces a spectral-domain augmentation strategy that explicitly targets underrepresented regions of the label space while preserving global graph structure and chemical validity. Unlike many existing approaches that rely on pseudo-labeling or sacrifice accuracy in well-represented regions, SPECTRA generates new, chemically coherent samples where data are sparse, mitigating imbalance without degrading overall performance. By combining spectral alignment with rare-target-aware sampling and validity-preserving reconstruction, it enables interpretable molecule generation and improves regression accuracy in rare but scientifically critical regimes.

## 3 METHOD

We propose a spectral, geometry-aware augmentation and learning pipeline for molecular property prediction that (i) constructs multi-attribute Laplacian representation from molecular graphs; (ii) aligns laplacians and nodes representations of different graphs using (Fused) Gromov–Wasserstein (FGW) couplings; (iii) interpolates eigenvalues and eigenvectors, along to node features in a stable orthonormal basis; and (iv) trains a spectral GNN with edge-aware Chebyshev convolutions on original and augmented samples. Figure 2 summarizes the workflow.

### 3.1 FROM SMILES TO MULTI-ATTRIBUTE GRAPHS

Given a SMILES string  $s$ , we construct the graph  $G = (V, E, \mathbf{X}, \mathbf{E}, y)$  with RDKit<sup>1</sup>. Nodes  $v \in V$  carry atom features  $\mathbf{X} \in \mathbb{R}^{n \times d}$  (OGB utilities), and each undirected edge  $(u, v) \in E$  has a 3D attribute vector (bond type, stereo, conjugation). We treat the three edge channels as separate weighted adjacencies  $\{\mathbf{W}^{(f)}\}_{f=1}^F$  ( $F=3$ ), and compute one (unnormalized) Laplacian per channel,

$$\mathbf{L}^{(f)} = \mathbf{D}^{(f)} - \mathbf{W}^{(f)}, \quad \mathbf{D}^{(f)} = \text{diag}(\mathbf{W}^{(f)} \mathbf{1}).$$

### 3.2 GEOMETRY-AWARE GRAPH MATCHING (FGW)

To establish node correspondence between two molecules  $A$  and  $B$ , we solve a Gromov–Wasserstein (GW) or Fused Gromov–Wasserstein (FGW) optimal transport problem on their zero-padded adjacency matrices  $\tilde{\mathbf{A}}, \tilde{\mathbf{B}} \in \mathbb{R}^{n \times n}$  (padding each graph to the larger node count). We define probability distributions  $p, q \in \Delta^n$  over the nodes of  $A$  and  $B$ , respectively. Each entry  $p_i$  (or  $q_j$ ) represents the relative “mass” assigned to node  $i$  in  $A$  (or node  $j$  in  $B$ ); in this work we use uniform weights so that  $p_i = 1/|V_A|$  and  $q_j = 1/|V_B|$ . The transport plan  $\mathbf{T} \in \mathbb{R}^{n \times n}$  then specifies how this probability mass is moved from each node of  $A$  to each node of  $B$ , effectively giving a soft alignment between their nodes. When node attributes are available, we use FGW with a cost matrix  $\mathbf{M}$  that measures feature dissimilarity; otherwise we use pure GW:

$$\mathbf{T}^* = \arg \min_{\mathbf{T} \in \Pi(p, q)} (1 - \alpha) \mathcal{L}_{\text{GW}}(\tilde{\mathbf{A}}, \tilde{\mathbf{B}}, \mathbf{T}) + \alpha \langle \mathbf{M}, \mathbf{T} \rangle,$$

where  $\Pi(p, q)$  is the set of couplings with marginals  $p$  and  $q$ ,  $\mathcal{L}_{\text{GW}}$  is the squared-loss GW discrepancy, and  $\alpha \in [0, 1]$  balances structural versus feature similarity. The resulting optimal coupling  $\mathbf{T}^*$  provides a soft node-to-node correspondence; we convert it into a hard one-to-one mapping using the Hungarian assignment on  $-\mathbf{T}^*$  and reorder  $\tilde{\mathbf{B}}$  and its features accordingly.

### 3.3 SPECTRAL ALIGNMENT AND INTERPOLATION

Given the matched pair, we diagonalize

$$\mathbf{L}_A = \mathbf{U}_A \boldsymbol{\Lambda}_A \mathbf{U}_A^\top, \quad \mathbf{L}_B = \mathbf{U}_B \boldsymbol{\Lambda}_B \mathbf{U}_B^\top,$$

and align eigenvector signs and bases with an orthogonal Procrustes map  $\mathbf{R}^* = \arg \min_{\mathbf{R} \in O(k)} \|\mathbf{U}_A^\top \mathbf{U}_B - \mathbf{R}\|_F$ , yielding  $\tilde{\mathbf{U}}_B = \mathbf{U}_B \mathbf{R}^*$ . We then interpolate eigenvalues and bases with a mixing coefficient  $\alpha \in (0, 1)$ :

$$\boldsymbol{\Lambda}_\alpha = (1 - \alpha) \boldsymbol{\Lambda}_A + \alpha \boldsymbol{\Lambda}_B, \quad \hat{\mathbf{U}} = (1 - \alpha) \mathbf{U}_A + \alpha \tilde{\mathbf{U}}_B, \quad \mathbf{U}_\alpha = \text{qr}(\hat{\mathbf{U}}),$$

and synthesize an intermediate Laplacian

$$\mathbf{L}_\alpha = \mathbf{U}_\alpha \boldsymbol{\Lambda}_\alpha \mathbf{U}_\alpha^\top$$

We repeat this per edge channel ( $F=3$ ).

**Node feature interpolation.** In the matched node domain we perform linear interpolation in the original node space:

$$\mathbf{X}_\alpha = (1 - \alpha) \mathbf{X}_A + \alpha \tilde{\mathbf{X}}_B,$$

where  $\tilde{\mathbf{X}}_B$  is  $\mathbf{X}_B$  permuted by the GW/FGW correspondence.

### 3.4 GRAPH RECONSTRUCTION FROM SPECTRA

For each channel we map  $\mathbf{L}_\alpha^{(f)}$  back to a nonnegative adjacency by removing degrees and clipping negatives,

$$\mathbf{W}_\alpha^{(f)} = \max(0, -\mathbf{L}_\alpha^{(f)} + \text{diag}(\mathbf{L}_\alpha^{(f)})), \quad \text{diag}(\mathbf{W}_\alpha^{(f)}) = \mathbf{0}.$$

We then assemble multi-attribute edges by scanning  $(u, v)$  with any positive channel weight and stacking per-channel features. The scalar label is interpolated as  $y_\alpha = (1 - \alpha)y_A + \alpha y_B$ .

<sup>1</sup>RDKit: <https://www.rdkit.org>

### 3.5 RARITY-AWARE PAIR SELECTION AND AUGMENTATION BUDGET

We compute a KDE over training labels to estimate density  $\rho(y)$  and define rarity weights  $w_i \propto 1/\rho(y_i)$  (normalized). Each training molecule  $i$  receives an augmentation budget  $\lfloor w_i \cdot N \cdot \text{perc} \rfloor$  ( $\text{perc} \in [0, 1]$  is a global rate). For molecule  $i$ , we sort neighbors by  $|y_i - y_j|$  and generate pairs  $(i, j)$  in that order, producing up to the allocated number of augmented graphs.

### 3.6 GRAPH VALIDITY AND CONVERSION BACK TO MOLECULES

To verify that the augmented graphs correspond to real chemical compounds, we convert each generated graph back into a SMILES string and validate its chemical consistency with RDKit. Each node’s feature vector is decoded into an atom specification (atomic number, charge, chirality, hybridization, aromaticity), falling back to reasonable defaults when attributes are missing. Bonds are reconstructed from edge attributes, including type, stereochemistry, and conjugation, while avoiding duplicates to maintain a valid simple graph.

The resulting editable molecule is then sanitized with RDKit to enforce valence rules, aromaticity perception, and proper connectivity. If strict sanitization fails, a relaxed mode attempts to correct hydrogen counts and minor inconsistencies. Finally, the molecule is converted to a canonical SMILES string; graphs that cannot be sanitized are marked invalid. The *validity* metric (Table 1) is defined as the fraction of generated graphs successfully converted to valid SMILES, ensuring that augmented samples are chemically meaningful rather than arbitrary graph structures.

### 3.7 SPECTRAL GNN WITH EDGE-AWARE CHEBYSHEV CONVOLUTIONS

We adopt a stack of  $L$  spectral blocks with ChebConv (Defferrard et al., 2016) and batch normalization:

$$\mathbf{H}^{(\ell+1)} = \text{Drop}\left(\text{SiLU}\left(\text{BN}(\text{ChebConv}_K(\mathbf{H}^{(\ell)}, \mathbf{A}, \mathbf{w}_e))\right)\right).$$

Multi-attribute edge features  $\mathbf{e}_{uv} \in \mathbb{R}^3$  are projected.

## 4 RESULTS

We evaluate our methods across three benchmark datasets (FreeSolv, ESOL, and Lipo), in detail in Appendix A.1. To assess the model’s ability to generate a diverse set of real molecules distinct from the training data (RQ1) we consider: quantitative estimate of drug-likeness (QED) (Bickerton et al., 2012), synthetic accessibility score (SA) (Ertl & Schuffenhauer, 2009), octanol–water partition coefficient (LogP) (Wildman & Crippen, 1999), exact molecular weight (MW), Bertz complexity (BCT) (Bertz, 1981), natural product likeness (NP) (Ertl et al., 2008). We also use standard metrics like validity, uniqueness, novelty.

We further evaluate the predictive performance of our model against state-of-the-art methods (RQ2) and analyze its behavior across the entire target domain to understand improvements in low-density regions compared to high-density regions (RQ3). We also assess the computational efficiency of our approach relative to existing methods (RQ4). Finally, we perform ablation studies to examine the impact of key design choices, with detailed results provided in Appendix A.2.

### 4.1 MOLECULE GENERATION QUALITY (RQ1)

To evaluate the quality of the generated molecules, we first assess their *validity*, *uniqueness*, and *novelty*. Validity is the fraction of generated molecules that are chemically valid, uniqueness is the fraction of valid molecules that are non-duplicate, and novelty is the fraction of unique molecules not present in the training set (Flam-Shepherd et al., 2022). Table 1 shows that all generated molecules are chemically valid (100% validity) and achieve high uniqueness and novelty. Uniqueness is slightly lower for FreeSolv, likely due to its smaller chemical space, but novelty remains consistently high, indicating that our augmentation strategy produces new, valid structures rather than replicating training molecules.

To further understand the impact on chemical space, we visualize original and augmented molecules using t-SNE on Morgan fingerprints (Figure 3). Augmented molecules populate sparse regions, im-

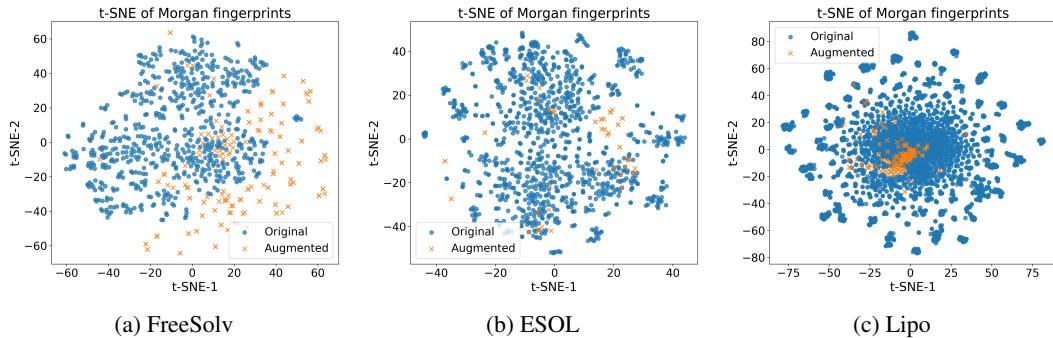


Table 3: **t-SNE visualization of Morgan fingerprints** comparing original and augmented samples for FreeSolv, ESOL, and Lipo.

proving coverage and mitigating distributional imbalance. This broader coverage helps reduce bias toward overrepresented regions and supports better generalization to underrepresented subspaces.

<b>Dataset</b>	<b>Validity</b>	<b>Uniqueness</b>	<b>Novelty</b>
FreeSolv	1.000	0.568	1.000
ESOL	1.000	0.661	0.949
Lipo	1.000	0.706	0.992

Table 1: Validity, Uniqueness, and Novelty of generated molecules across datasets.

Statistic	FreeSolv		ESOL		Lipo	
	Orig	Aug	Orig	Aug	Orig	Aug
Atoms <sub>min</sub>	1	5	1	2	7	10
Atoms <sub>mean</sub>	8.73	12.06	13.28	19.83	27.04	22.80
Atoms <sub>max</sub>	24	20	55	28	115	115
Rings <sub>min</sub>	0	0	0	0	0	0
Rings <sub>mean</sub>	0.66	1.77	1.39	2.86	3.49	3.25
Rings <sub>max</sub>	5	8	8	7	13	9

Table 2: Atom and ring statistics for Original vs. Augmented molecules across datasets.

We also examine structural complexity by comparing atom and ring counts between original and augmented molecules (Table 2). Augmented molecules are generally larger and more cyclic, with increased mean atom and ring counts in FreeSolv and ESOL and comparable ranges in Lipo. This shows that our method expands scaffold diversity while staying within chemically reasonable bounds, complementing the improved coverage and property-target alignment.

Finally, we assess whether augmented molecules preserve relationships between task targets and key molecular properties. Figure 3 shows the joint distributions of five properties (LogP, SA, QED, MW, and BT) versus the task targets. Across all datasets, augmented molecules (orange crosses) follow trends similar to the originals (blue circles). Notably, augmented samples extend property information, with some of them respecting target-property correlation, improving coverage while maintaining realistic property distributions. Overall, our augmentation method generates valid, novel, and structurally diverse molecules that respect domain-relevant and property dependencies while expand the chemical space in a task-aligned way.

#### 4.2 PREDICTIVE PERFORMANCE (RQ2)

We compare our proposed method (**SPECTRA**) against representative state-of-the-art molecular representation learning models, including contrastive methods (GraphCL(You et al., 2020), MolCLR (Wang et al., 2022)), language–graph hybrids (Molformer Ross et al. (2022), Chemb Defferrard et al. (2016)), and recent GNN-based frameworks (HiMol (Zang et al., 2023), SGIR (Liu et al., 2023a)). Table 4 reports both the general prediction accuracy, measured by mean absolute error (MAE), and performance under imbalance, measured by squared error relevance area (SERA). The experimental setup is presented in Appendix A.3.

SPECTRA achieves consistently strong performance across all datasets. On ESOL and Lipo, SPECTRA is competitive with the best-performing models, and on FreeSolv it surpasses most baselines with the second-best overall performance. While SGIR attains the best MAE and SERA scores on average, SPECTRA achieves a stable balance across datasets and excels in capturing underrep-

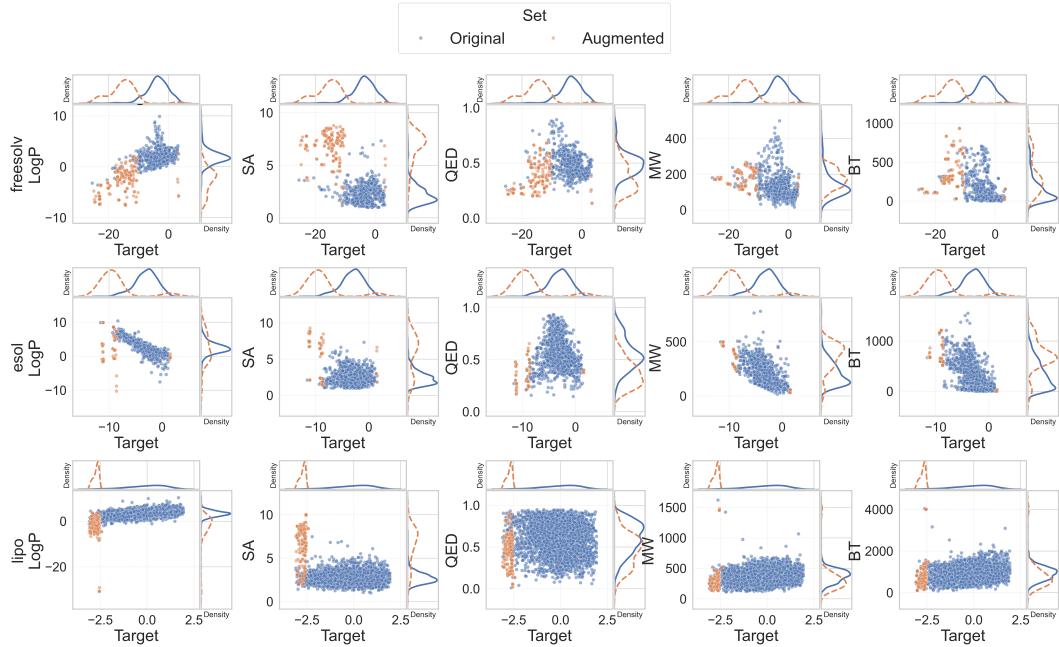


Figure 3: Joint distribution plots of molecular properties versus task targets for original (blue, circles, solid marginals) and augmented (orange, crosses, dashed marginals) molecules. Each row corresponds to a dataset (FreeSolv, ESOL, Lipo), and each column shows one computed property (LogP, SA, QED, MW, BT).

Model	MAE (mean $\pm$ var)			SERA (mean $\pm$ var)		
	ESOL	FreeSolv	Lipo	ESOL	FreeSolv	Lipo
Chemb	$0.59 \pm 0.32$	$0.93 \pm 1.26$	$0.41 \pm 0.15$	$0.25 \pm 0.01$	$1.07 \pm 0.19$	$0.11 \pm 0.00$
GraphCL	$0.78 \pm 0.40$	$1.76 \pm 2.30$	$0.73 \pm 0.30$	$0.36 \pm 0.00$	$2.59 \pm 0.45$	$0.40 \pm 0.00$
HiMol	$0.51 \pm 0.22$	$0.97 \pm 1.46$	$0.41 \pm 0.14$	$0.17 \pm 0.00$	$1.34 \pm 0.74$	$0.10 \pm 0.00$
MolCLR	$0.73 \pm 0.40$	$1.12 \pm 1.31$	$0.43 \pm 0.14$	$0.32 \pm 0.00$	$1.36 \pm 0.18$	$0.11 \pm 0.00$
Molformer	$1.66 \pm 1.68$	$2.84 \pm 6.87$	$0.81 \pm 0.38$	$2.77 \pm 0.01$	$10.61 \pm 12.16$	$0.54 \pm 0.00$
SGIR	<b><math>0.46 \pm 0.19</math></b>	<b><math>0.68 \pm 0.85</math></b>	<b><math>0.37 \pm 0.13</math></b>	<b><math>0.13 \pm 0.00</math></b>	<b><math>0.69 \pm 0.05</math></b>	<b><math>0.09 \pm 0.00</math></b>
SPECTRA	<u><math>0.53 \pm 0.28</math></u>	<u><math>0.77 \pm 1.05</math></u>	<u><math>0.38 \pm 0.13</math></u>	<u><math>0.20 \pm 0.00</math></u>	<u><math>0.95 \pm 0.29</math></u>	<u><math>0.09 \pm 0.00</math></u>

Table 4: Mean absolute error (MAE) and SERA with variance for each model across three datasets. Lower values indicate better performance. Bold models are the best results while, underlined are the second best.

resented regions, as reflected in its low SERA values. This indicates that our augmentation and spectral alignment strategies effectively improve prediction in imbalanced regimes without sacrificing overall performance.

#### 4.3 ERROR DISTRIBUTION ACROSS TARGET RANGES (RQ3)

Figure 4 further dissects MAE by target value ranges. We observe that baseline models often suffer from considerably higher errors in the low-density regions, consistent with the imbalance in the training data. In contrast, SPECTRA demonstrates markedly lower errors in these sparse regions, highlighting its strength in addressing imbalance.

#### 4.4 EFFICIENCY AND PARETO OPTIMALITY (RQ4)

Besides accuracy, computational efficiency is a key consideration in real-world applications. Figure 5 illustrates the trade-off between runtime and accuracy for all models, with Pareto frontiers

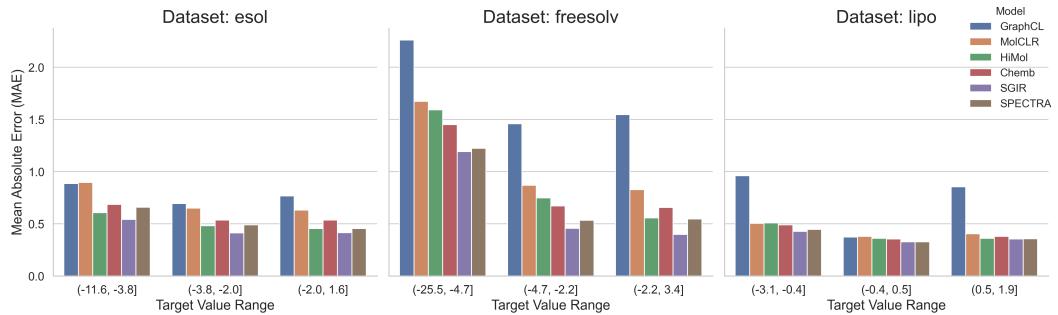


Figure 4: **Mean Absolute Error (MAE) distribution** across target value ranges for each dataset. Colors correspond to different models as indicated in the legend.

identified for each dataset. SPECTRA consistently lies on or very close to the Pareto frontier, indicating that it achieves a favorable trade-off between performance and efficiency. Compared to transformer-based models such as Molformer, which incur substantial runtime costs, SPECTRA achieves competitive or superior accuracy with significantly reduced computational overhead.

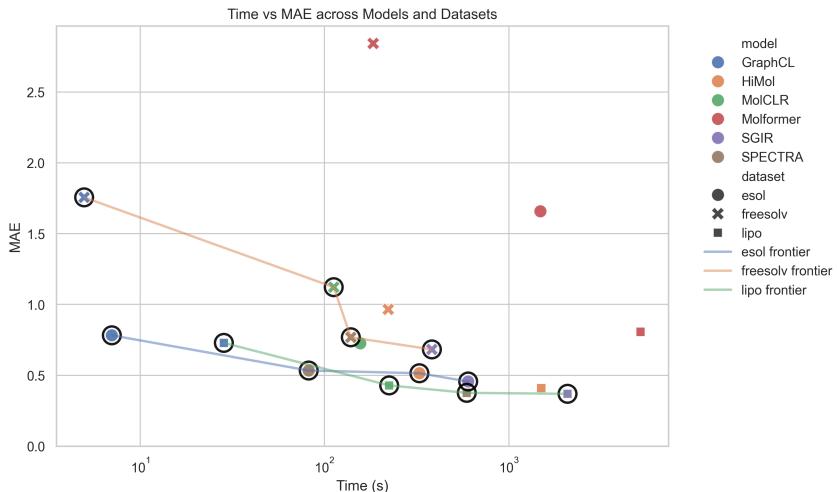


Figure 5: **Time vs. MAE across models and datasets.** Each point represents the average runtime (log scale) and mean absolute error (MAE) of a model–dataset pair. Black hollow circles and connecting lines indicate the Pareto frontier for each dataset.

## 5 CONCLUSION

Experiments across benchmark datasets show that our method improves predictive accuracy in rare but critical regimes, preserves property–target correlations, and achieves a favorable balance between accuracy and efficiency. These results establish spectral augmentation as a promising and interpretable strategy for tackling imbalance in molecular property prediction and related graph-structured scientific domains. Future work will explore extending the framework to multi-property prediction as well as incorporating additional modalities such as 3D features.

## REFERENCES

- Rafael Lopes Almeida, Vinícius Gonçalves Matarollo, and Frederico Gualberto Ferreira Coelho. Overcoming class imbalance in drug discovery problems: Graph neural networks and balancing approaches. *Journal of Molecular Graphics and Modelling*, 126:108627, 2024.

- Steven H Bertz. The first general index of molecular complexity. *Journal of the American Chemical Society*, 103(12):3599–3601, 1981.
- G Richard Bickerton, Gaia V Paolini, Jérémie Besnard, Sorel Muresan, and Andrew L Hopkins. Quantifying the chemical beauty of drugs. *Nature chemistry*, 4(2):90–98, 2012.
- Deyu Bo, Chuan Shi, Lele Wang, and Renjie Liao. Specformer: Spectral graph neural networks meet transformers. *arXiv preprint arXiv:2303.01028*, 2023a.
- Deyu Bo, Chuan Zheng, Xincheng Wang, Peipei Jiao, Shirui Zhou, Hao Zhang, Zhewei Wei, and Chuan Shi. A survey on spectral graph neural networks. *arXiv preprint arXiv:2302.05631*, 2023b.
- Paula Branco, Luís Torgo, and Rita P Ribeiro. 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.
- Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019.
- Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- Yangyang Chen, Zixu Wang, Lei Wang, Jianmin Wang, Pengyong Li, Dongsheng Cao, Xiangxiang Zeng, Xiucai Ye, and Tetsuya Sakurai. Deep generative model for drug design from protein target sequence. *Journal of Cheminformatics*, 15(1):38, 2023.
- Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.
- Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 29, 2016.
- Peter Ertl and Ansgar Schuffenhauer. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *Journal of cheminformatics*, 1(1):8, 2009.
- Peter Ertl, Silvio Roggo, and Ansgar Schuffenhauer. Natural product-likeness score and its application for prioritization of compound libraries. *Journal of chemical information and modeling*, 48(1):68–74, 2008.
- Zhe Fan, Junda Yu, Xiangyu Zhang, Yuhua Chen, Shuqian Sun, Yuyang Zhang, Ming Chen, Feng Xiao, Wei Wu, Xiang-Nan Li, et al. Reducing overconfident errors in molecular property classification using posterior network. *Patterns*, 2024.
- Daniel Flam-Shepherd, Kevin Zhu, and Alán Aspuru-Guzik. Language models can learn complex molecular distributions. *Nature Communications*, 13(1):3293, 2022.
- Yu Gong, Greg Mori, and Frederick Tung. Ranksim: Ranking similarity regularization for deep imbalanced regression. *arXiv preprint arXiv:2205.15236*, 2022.
- Woosung Jeon and Dongsup Kim. Autonomous molecule generation using reinforcement learning and docking to develop potential novel inhibitors. *Scientific reports*, 10(1):22104, 2020.
- Wengong Jin, Regina Barzilay, and Tommi Jaakkola. Junction tree variational autoencoder for molecular graph generation. *International Conference on Machine Learning*, pp. 2323–2332, 2018.
- Mohammadreza Karamad, Rishi Magar, Yanming Shi, Samira Siahrostami, Ian D Gates, and Amir Barati Farimani. Orbital graph convolutional neural network for material property prediction. *Physical Review Materials*, 4(9):093801, 2020.

- Yash Khemchandani, Stephen O’Hagan, Soumitra Samanta, Neil Swainston, Timothy J Roberts, Danushka Bollegala, and Douglas B Kell. Deepgraphmolgen, a multi-objective, computational strategy for generating molecules with desirable properties: a graph convolution and reinforcement learning approach. *Journal of cheminformatics*, 12(1):53, 2020.
- Myeonghun Lee and Kyoungmin Min. Mgcvae: multi-objective inverse design via molecular graph conditional variational autoencoder. *Journal of chemical information and modeling*, 62(12):2943–2950, 2022.
- Tianyi Li, Hongxu Yin, Chuan Shi, and Wei Lin. Large-scale spectral graph neural networks via laplacian sparsification: Technical report. *arXiv preprint arXiv:2501.04570*, 2025.
- Jaechang Lim, Seongok Ryu, Kyubyong Park, Yo Jun Choe, Jiyeon Ham, and Woo Youn Kim. 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.
- Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pp. 2980–2988, 2017.
- Gang Liu, Tong Zhao, Eric Inae, Tengfei Luo, and Meng Jiang. Semi-supervised graph imbalanced regression. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1453–1465, 2023a.
- Gang Liu, Tong Zhao, Eric Inae, Tengfei Luo, and Meng Jiang. 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, 2023b. Association for Computing Machinery. ISBN 9798400701030. doi: 10.1145/3580305.3599497. URL <https://doi.org/10.1145/3580305.3599497>.
- Aditya Krishna Menon, Sadeep Jayasumana, Ankit Singh Rawat, Himanshu Jain, Andreas Veit, and Sanjiv Kumar. Long-tail learning via logit adjustment. *arXiv preprint arXiv:2007.07314*, 2020.
- Oleksii Prykhodko, Simon Viet Johansson, Panagiotis-Christos Kotsias, Josep Arús-Pous, Esben Jannik Bjerrum, Ola Engkvist, and Hongming Chen. A de novo molecular generation method using latent vector based generative adversarial network. *Journal of cheminformatics*, 11(1):74, 2019.
- Jiawei Ren, Mingyuan Zhang, Cunjun Yu, and Ziwei Liu. Balanced mse for imbalanced visual regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7926–7935, 2022.
- Rita P. Ribeiro and Nuno Moniz. Imbalanced regression and extreme value prediction. *Machine Learning*, 109(9):1803–1835, 2020a.
- Rita P. Ribeiro and Nuno Moniz. Imbalanced regression and extreme value prediction. *Machine Learning*, 109(9):1803–1835, September 2020b. ISSN 1573-0565. doi: 10.1007/s10994-020-05900-9. URL <https://doi.org/10.1007/s10994-020-05900-9>.
- Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkitt Padhi, Youssef Mrroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022.
- Junjiao Tian, Yen-Cheng Liu, Nathaniel Glaser, Yen-Chang Hsu, and Zsolt Kira. Posterior re-calibration for imbalanced datasets. *Advances in neural information processing systems*, 33: 8101–8113, 2020.
- Jessica Vamathevan, Dominic Clark, Paul Czodrowski, Ian Dunham, Edgardo Ferran, George Lee, Bin Li, Anant Madabhushi, Parantu Shah, Michaela Spitzer, et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6):463–477, 2019.
- Xiyuan Wang and Ming Zhang. How powerful are spectral graph neural networks. *arXiv preprint arXiv:2205.11172*, 2022.

- Yuyang Wang, Jianren Wang, Zhonglin Cao, and Amir Barati Farimani. Molecular contrastive learning of representations via graph neural networks. *Nature Machine Intelligence*, 4(3):279–287, 2022.
- Scott A Wildman and Gordon M Crippen. Prediction of physicochemical parameters by atomic contributions. *Journal of chemical information and computer sciences*, 39(5):868–873, 1999.
- Zhenqin Wu, Bharath Ramsundar, Evan N Feinberg, Joseph Gomes, Caleb Geniesse, Aneesh S Pappu, Karl Leswing, and Vijay Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical science*, 9(2):513–530, 2018.
- Ruoyan Xia, Chao Zhang, and Yongdong Zhang. A novel graph oversampling framework for node classification in class-imbalanced graphs. *Science China Information Sciences*, 67(1):162101, 2024.
- Zhaoping Xiong, Dingyan Wang, Xiaohong Liu, Feisheng Zhong, Xutong Wan, Xiang Li, Zhaojian Li, Xiaomin Luo, Kaixian Chen, Hualiang Jiang, 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.
- Kaiqi Yang, Haoyu Han, Wei Jin, and Hui Liu. Spectral-aware augmentation for enhanced graph representation learning. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 2837–2847, 2024.
- Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. Delving into deep imbalanced regression. In *International conference on machine learning*, pp. 11842–11851. PMLR, 2021.
- Rufan Yao, Zhenhua Shen, Xinyi Xu, Guixia Ling, Rongwu Xiang, Tingyan Song, Fei Zhai, and Yuxuan Zhai. Knowledge mapping of graph neural networks for drug discovery: a bibliometric and visualized analysis. *Frontiers in Pharmacology*, 15, 2024.
- Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823, 2020.
- Xuan Zang, Xianbing Zhao, and Buzhou Tang. Hierarchical molecular graph self-supervised learning for property prediction. *Communications Chemistry*, 6(1):34, 2023.
- Bin Zhang and Mengjun Tu. A review on graph neural networks for predicting synergistic drug combinations. *Artificial Intelligence Review*, 2023.
- Nannan Zong, Songzhi Su, and Changle Zhou. Boosting semi-supervised learning under imbalanced regression via pseudo-labeling. *Concurrency and Computation: Practice and Experience*, 36(19): e8103, 2024.

## A APPENDIX

### A.1 DATASET DETAILS

Our experimental evaluation uses molecular regression tasks from MoleculeNet (Wu et al., 2018), specifically ESOL, FreeSolv, and Lipophilicity (Lipo). A brief summary of these datasets is provided in Table 5.

### A.2 ABLATION STUDY

To disentangle the individual contributions of our design choices, we conduct an ablation study on spectral alignment (FGW) and KDE-based augmentation. Table 6 reports the results across ESOL, FreeSolv, and Lipo. Excluding FGW alignment leads to a marked performance drop on FreeSolv, highlighting the necessity of geometry-aware alignment when handling structurally diverse molecules. Similarly, removing KDE-based augmentation degrades performance in imbalanced regions, demonstrating the role of density-aware augmentation in enhancing generalization. Overall, the full model consistently achieves the best or near-best results across datasets.

Table 5: Summary of Molecular Property Datasets

Dataset	# of Compounds	Description
ESOL	1,128	Water solubility (log solubility in mol/L)
FreeSolv	642	Hydration free energy in water
Lipophilicity	4,200	Octanol/water distribution coefficient (logD at pH 7.4)

Table 6: Ablation study with incremental addition of augmentation (Aug), alignment (Align), and KDE prior. Results are reported as mean (std) of per-sample errors over multiple runs. Best results per dataset are highlighted in bold.

Aug	Align	KDE	ESOL	FreeSolv	Lipo
✗	✗	✗	0.586 (0.568)	0.926 (1.125)	0.408 (0.384)
✓	✗	✗	0.552 (0.562)	0.863 (1.128)	0.384 (0.371)
✓	✗	✓	<b>0.534 (0.515)</b>	0.869 (1.113)	0.378 (0.362)
✓	✓	✓	0.534 (0.525)	<b>0.769 (1.023)</b>	<b>0.377 (0.359)</b>

### A.3 EXPERIMENTAL SETUP

All experiments were conducted on a Linux server equipped with two 12-core Intel(R) Haswell processors, 256 GB of RAM, and four NVIDIA A100 GPUs, each with 80 GB of memory. Our method is implemented in Python 3.8.19 using PyTorch 2.1.2. We used Chebyshev GCN (cheb) (Defferrard et al., 2016). We perform a manual hyperparameter search over the following ranges:

- **Hidden dimension:** {128, 256, 512}
- **Number of layers:** {3, 4, 5}
- **Dropout:** {0.0, 0.1, 0.3}
- **Learning rate:** { $10^{-3}$ ,  $2 \times 10^{-3}$ ,  $5 \times 10^{-4}$ }
- **Chebyshev filter order ( $k$ ):** {2, 3, 5}
- **Epochs:** {500}
- **Batch size:** {32, 64}
- **Alpha:** {0.1, 0.2, 0.3, 0.4, 0.5}

Our code and data are available on GitHub<sup>2</sup>.

### B THE USE OF LARGE LANGUAGE MODELS (LLMs)

LLM was used just to polish grammar.

<sup>2</sup><https://anonymous.4open.science/r/SPECTRA-0D3C>