Are we Making Much Progress? Revisiting Chemical Reaction Yield Prediction from an Imbalanced Regression Perspective

Yihong Ma
University of Notre Dame
yma5@nd.edu

Xiaobao Huang
University of Notre Dame
xhuang2@nd.edu

Bozhao Nan
University of Notre Dame
bnan@nd.edu

Nuno Moniz
University of Notre Dame
nuno.moniz@nd.edu

Xiangliang Zhang
University of Notre Dame
xzhang33@nd.edu

Olaf Wiest
University of Notre Dame
owiest@nd.edu

Nitesh V. Chawla
University of Notre Dame
nchawla@nd.edu

ABSTRACT

The yield of a chemical reaction quantifies the percentage of the target product formed in relation to the reactants consumed during the chemical reaction. Accurate yield prediction can guide chemists toward selecting high-yield reactions during synthesis planning, offering valuable insights before dedicating time and resources to wet lab experiments. While recent advancements in yield prediction have led to overall performance improvement across the entire yield range, an open challenge remains in enhancing predictions for high-yield reactions, which are of greater concern to chemists. In this paper, we argue that *the performance gap in high-yield predictions results from the imbalanced distribution of real-world data skewed towards low-yield reactions, often due to unreacted starting materials and inherent ambiguities in the reaction processes.* Despite this data imbalance, existing yield prediction methods continue to treat different yield ranges equally, assuming a balanced training distribution. Through extensive experiments on three real-world yield prediction datasets, we emphasize the urgent need to reframe reaction yield prediction as an imbalanced regression problem. Finally, we demonstrate that incorporating simple cost-sensitive reweighting methods can significantly enhance the performance of yield prediction models on underrepresented high-yield regions.

CCS CONCEPTS

• **Applied computing** → **Chemistry**; • **Computing methodologies** → **Machine learning**.

KEYWORDS

Reaction yield prediction; data imbalance; regression tasks

ACM Reference Format:
Yihong Ma, Xiaobao Huang, Bozhao Nan, Nuno Moniz, Xiangliang Zhang, Olaf Wiest, and Nitesh V. Chawla. 2024. Are we Making Much Progress? Revisiting Chemical Reaction Yield Prediction from an Imbalanced Regression Perspective. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589335.3651470

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05.
https://doi.org/10.1145/3589335.3651470

1 INTRODUCTION

Recent advancements in machine learning have introduced a paradigm shift in the field of computational chemistry [5]. These breakthroughs have led to a diverse array of machine learning models that now play critical roles in assisting chemists across a broad spectrum of tasks, including but not limited to retrosynthesis, product prediction, and drug discovery. Among this multifaceted landscape, the prediction of reaction yields [1, 12, 13, 15, 17] emerges as an issue of paramount importance in the domain of synthesis planning, where complex molecules are synthesized through a sequence of reaction steps. Based on the empirical categorization, yields above 67% are classified as high yields and those below 33% are classified as low yields [17]. In this context, the occurrence of a low-yield reaction within this sequence can drastically impact the feasibility and overall efficiency of the synthesis process. As a result, chemists often prioritize the accurate prediction of high-yield reactions.

While the introduction of numerous yield prediction models has indeed showcased improved performance across the entire yield range, the challenge of effectively enhancing performance for high-yield reactions remains an open problem [9]. In real-world scenarios, yield data often exhibits a highly imbalanced distribution, with high yield values being much rarer than lower ones, despite their greater importance to chemists in synthesis planning. In this paper, we argue that *the increased difficulty in predicting high-yield reactions stems from its limited availability of data samples, often due to unreacted starting materials and inherent ambiguities in the reaction processes.* Despite the presence of such data imbalance, existing yield prediction methods continue to treat different yield ranges equally with the false assumption of a balanced data distribution.

To gain a deeper insight into the field's actual progress, we conduct extensive experiments to benchmark six state-of-the-art yield prediction methods on three real-world datasets. Surprisingly, the results become less impressive than claimed when we take data imbalance into account. We discover that *the overall good performance across the entire yield spectrum primarily results from*

enhancing performance in areas with sufficient data, typically the low-yield range, while overlooking the significant performance gap in underrepresented high-yield regions. This finding has motivated us to revisit reaction yield prediction and reformulate it as an imbalanced regression problem, a well-established topic in machine learning.

Unlike imbalanced classification, reaction yield prediction involves regression rather than classification, and there has been limited exploration of addressing data imbalance in the regression context [11]. Most prior research on imbalanced regression has directly adapted the SMOTE algorithm [3] to regression settings [2, 14]. However, the continuous nature of target labels in regression tasks makes these adaptations less practical. A more intuitive solution is to apply cost-sensitive re-weighting strategies [4, 7, 16] that can be seamlessly combined with various regression models. We demonstrate that incorporating these simple methods can significantly enhance the performance of existing yield prediction models on underrepresented high-yield regions without sacrificing the overall performance too much. We believe these findings have the potential to redirect the future research direction in reaction yield prediction, benefiting both chemistry and machine learning communities. In summary, the contributions of this paper include:

- We are the first to introduce the novel concept of reformulating reaction yield prediction as an imbalanced regression problem.
- We conduct comprehensive experiments on three real-world yield prediction datasets to uncover and understand the limitations of existing models when predicting high-yield reactions.
- We demonstrate that incorporating cost-sensitive re-weighting methods into existing yield prediction models can lead to significant performance improvements on high-yield reactions.

2 THE EXAMINATION OF EXISTING YIELD PREDICTION METHODS

In this section, we begin by introducing the definitions of reaction yield prediction and imbalanced regression. We then proceed to evaluate six yield prediction methods on three real-world datasets.

2.1 Preliminaries

Definition 1: Reaction yield prediction. Reaction yield prediction is a regression problem that predicts the yield value $y \in [0, 100]$ of a chemical reaction $rxn = (\mathcal{R}, P)$ composed of multiple reactant molecules $\mathcal{R}$ and a single product molecule P.

Definition 2: Imbalanced regression. Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N}$ denote the training dataset of a regression problem, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is the input feature vector and $y_i \in \mathbb{R}$ is the label. We divide the label space $\mathcal{Y}$ into K disjoint bins with equal intervals, i.e., $[b_0, b_1), [b_1, b_2), \ldots, [b_{K-1}, b_K)$. Let C_k be the set of data samples in the k-th bin with $k \in \{1, 2, \ldots, B\}$. The data imbalance occurs when the label distribution is highly skewed, i.e., $\frac{\max_k |C_k|}{\min_k |C_k|} \gg 1$.

2.2 Evaluation Settings for Yield Prediction

2.2.1 Datasets. We use three real-world datasets for predicting reaction yields, sourced from either high-throughput experimentation (HTE) or electronic laboratory notebooks (ELN). Following prior research on imbalanced regression [4, 8, 16], we categorize the bins within the target yield space into three disjoint subsets: *many-shot* (bins with over $\#_{\text{upper}}$ reactions), *medium-shot* (bins with

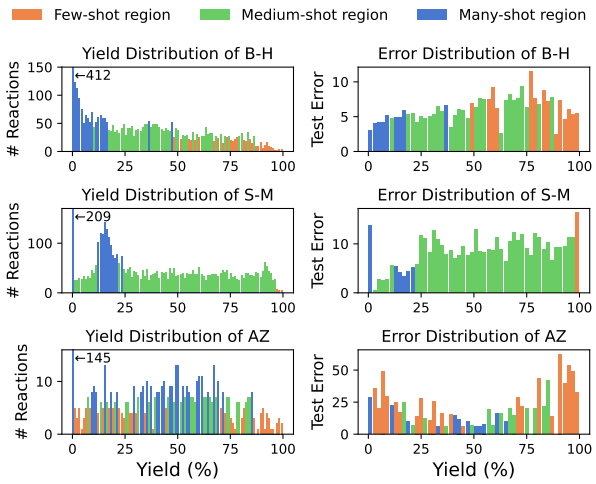

Figure 1: A comparison between yield distributions (left) and test error distributions (right) on three real-world datasets.

$\#_{\text{lower}}$ to $\#_{\text{upper}}$ reactions), and *few-shot* (bins with fewer than $\#_{\text{lower}}$ reactions) regions, based on their respective numbers of training samples. For all three datasets, we set the bin size $|b_k - b_{k-1}|$ to 1. The left section of Figure 1 visualizes the division of these regions.

- B-H [1]: It comprises 3,955 Buchward-Hartwig reactions from HTE, with the number of reactions per bin varying between 1 and 412. Here, $\#_{\text{lower}}$ is set to 25, and $\#_{\text{upper}}$ is set to 50.
- S-M [10]: It consists of 5,760 Suzuki-Miyaura reactions from HTE, with the number of reactions per bin ranging from 1 to 209. Here, $\#_{\text{lower}}$ is set to 20, and $\#_{\text{upper}}$ is set to 65.
- AZ [12]: It includes 750 Buchward-Hartwig reactions from ELN at AstraZeneca, with the number of reactions per bin ranging from 0 to 145. Here, $\#_{\text{lower}}$ is set to 3, and $\#_{\text{upper}}$ is set to 5.

2.2.2 Yield prediction methods.

- **Machine learning methods:** Random Forest (RF), XGBoost, Support Vector Machines (SVM);
- **Deep learning methods:** Multi-layer Perceptron (MLP), YieldGNN [12], Yield-BERT [13].

2.2.3 Evaluation pipeline. We report yield prediction results on *many-shot*, *medium-shot*, and *few-shot* regions as well as on the entire yield space (i.e., the *all* region). In all three datasets, 70% of the data is used for training and the remaining 30% is reserved for testing. Our evaluation employs common yield prediction metrics: mean absolute error (MAE) and root mean square error (RMSE). Additionally, we also utilize the geometric mean of L_1 errors (G-Mean) as a supplementary metric. Lower values (↓) of MAE, RMSE, and G-Mean indicate better yield prediction performance.

2.2.4 Implementation details. For RF, XGBoost, SVM, and MLP, the input features include structural fingerprints (e.g., ECFP), chemical properties (e.g., NMR shifts, HOMO/LUMO energies, vibrations, dipole moments), and reaction-specific parameters (e.g., scale, volume, temperature). For Yield-BERT, the SMILES string of the reaction is used as input; an encoder based on a pre-trained BERT [6] for SMILES is employed, with a k-Nearest Neighbors (kNN) regressor serving as the decoder. For YieldGNN, we construct graph structures to represent the molecules involved in the reaction; a

Dataset	Method	MAE ↓				RMSE ↓				G-Mean ↓			
		All	Many	Med.	Few	All	Many	Med.	Few	All	Many	Med.	Few
B-H	RF	5.5±0.2	4.4±0.2	6.0±0.2	6.7±0.6	8.1±0.3	7.4±0.3	8.4±0.4	9.1±0.8	2.8±0.1	1.8±0.1	3.6±0.2	4.1±0.4
	XGBoost	4.7±0.2	3.8±0.2	5.2±0.2	5.7±0.5	6.9±0.3	6.1±0.4	7.1±0.5	7.8±0.7	2.7±0.1	2.0±0.1	3.2±0.1	3.4±0.3
	SVM	14.6±0.3	12.9±0.5	12.9±0.7	26.1±1.1	18.5±0.4	15.8±0.5	17.2±0.8	28.6±0.9	9.2±0.3	8.7±0.5	7.7±0.6	22.2±1.6
	MLP	4.5±0.2	3.4±0.3	5.3±0.2	5.3±0.4	7.0±0.5	6.0±1.0	7.6±0.6	7.3±0.6	2.5±0.1	1.8±0.1	3.1±0.1	3.3±0.3
	YieldGNN	8.7±7.9	8.6±9.5	7.6±4.6	12.8±14.5	11.1±8.7	10.6±9.1	10.1±5.9	14.7±14.7	3.0±0.4	2.3±0.5	3.5±0.4	3.9±0.5
	Yield-BERT	5.5±0.3	3.8±0.3	6.6±0.4	6.6±0.5	8.4±0.4	7.1 ±1.4	9.1±0.6	8.9±0.8	3.2±0.1	2.2±0.1	4.1±0.4	4.1±0.5
Avg. Ranking		-	**1.2**	**1.8**	**3.0**	-	**1.2**	**2.2**	**2.7**	-	**1.2**	**1.8**	**3.0**
S-M	RF	8.0±0.2	6.1±0.4	8.8±0.3	12.8±6.1	11.8±0.4	11.2±0.9	12.0±0.3	15.9±7.6	4.0±0.2	2.3±0.1	5.1±0.3	9.8±3.8
	XGBoost	7.2±0.2	6.1±0.4	7.7±0.2	6.8±4.7	10.5±0.2	10.1±0.8	10.7±0.2	9.7±6.3	4.1±0.1	3.1±0.2	4.6±0.1	3.9±2.7
	SVM	16.5±0.2	14.8±0.5	17.1±0.2	31.6±6.0	20.3±0.2	18.7±0.6	20.8±0.3	33.0±6.1	11.1±0.2	9.6±0.3	11.7±0.3	30.3±6.0
	MLP	7.5±0.3	6.1±0.5	8.1±0.4	8.5±6.8	11.1±0.4	10.7±0.9	11.2±0.5	12.3±9.3	4.1±0.2	2.9±0.2	4.8±0.3	4.5±3.5
	YieldGNN	9.1±1.2	7.2±0.9	9.9±1.9	10.8±7.3	12.6±1.5	10.8±1.4	13.2±2.2	13.5±8.2	5.4±0.6	4.1±0.5	6.1±1.2	8.0±6.9
	Yield-BERT	9.1±0.4	6.3±0.4	10.2±0.5	5.1±5.4	13.3±0.6	11.2±0.9	14.1±6.3	7.7±7.6	5.0±0.2	3.3±0.2	6.0±0.3	2.3±2.7
Avg. Ranking		-	**1.2**	**2.3**	**2.5**	-	**1.3**	**2.3**	**2.3**	-	**1.2**	**2.5**	**2.3**
AZ	RF	20.3±0.8	19.1±0.7	23.6±2.5	32.0±4.3	25.2±0.9	23.6±0.8	29.8±3.0	36.4±3.9	13.8±0.8	13.1±0.6	16.0±3.0	26.0±6.4
	XGBoost	20.6±1.0	19.5±1.0	23.9±3.4	30.4±6.1	27.2±1.3	25.6±1.3	31.9±3.6	37.9±6.0	11.8±1.0	11.3±1.1	14.0±3.6	19.7±6.5
	SVM	25.0±0.8	23.5±0.7	28.0±2.4	41.7±3.9	28.9±0.9	27.1±0.7	32.5±2.4	43.9±3.8	18.6±0.9	17.5±0.7	20.8±2.9	37.6±5.8
	MLP	22.1±1.7	21.5±1.3	25.0±5.0	26.0±6.9	29.7±2.1	28.9±1.9	33.4±4.9	32.8±7.4	12.0±1.2	11.6±0.9	14.6±5.0	15.6±6.7
	YieldGNN	22.5±0.8	21.5±0.6	25.5±3.3	33.4±6.7	28.0±1.0	26.5±1.0	32.0±3.7	38.6±6.0	15.2±1.1	14.6±0.9	17.0±3.1	25.7±9.2
	Yield-BERT	25.6±2.5	24.3±2.6	28.7±3.8	38.4±10.1	32.5±3.0	30.7±3.0	37.5±3.6	45.2±9.9	15.7±1.5	15.1±1.7	18.0±4.7	26.2±10.8
Avg. Ranking		-	**1.0**	**2.0**	**3.0**	-	**1.0**	**2.2**	**2.8**	-	**1.0**	**2.0**	**3.0**

Table 1: Reaction yield prediction results on three real-world datasets. We report the average performance across 10 repetitions in all experiments and present the average rankings for the *many-shot*, *medium-shot*, and *few-shot* regions, respectively.

GNN is used to encode the reaction, while an MLP is used to decode the reaction embedding into yield predictions. We employ the L_1 distance as the training loss $\mathcal{L}$ in all experiments.

2.3 Uncovering and Understanding the Performance Gap in High-Yield Predictions

Figure 1 provides an insightful comparison between yield distributions and test error distributions, both as functions of yield values. Regarding the yield distribution, we note that the *few-shot* region predominantly comprises high-yield reactions, while the *many-shot* region primarily consists of low-yield reactions. Specifically in the B-H and S-M datasets, 81% and 100% of the *few-shot* reactions fall into the high-yield category, while 89% and 100% of the *many-shot* reactions belong to the low-yield category. This finding establishes a clear connection between yield values and their distributions.

To quantify the impact of data imbalance on prediction errors, we compute the Pearson correlation coefficients between the testing error distribution and the training yield distribution. Across all three reaction yield prediction datasets, we consistently observe negative correlation coefficients, with values of -0.42, -0.28, and -0.07, respectively. Moreover, in the right part of Figure 1, it is evident that the *few-shot* region (in orange) exhibits the largest test error, while the *many-shot* region (in blue) demonstrates the smallest error. To complement this observation, we further evaluate the yield prediction performance of six state-of-the-art models using three metrics and report the average performance rankings for the three regions in Table 1. As a result, the average ranking indicates a decline in model performance as we transition from the *many-shot* region to the *medium-shot* and *few-shot* regions.

Therefore, both observations from Figure 1 and Table 1 converge to the same conclusion: During the training process, low-yield values with a larger number of data samples tend to be learned better in comparison to those high yields with fewer samples. This highlights the demand for specialized machine learning techniques that can effectively address the challenge of data imbalance.

3 MITIGATING DATA IMBALANCE FOR BETTER HIGH-YIELD PREDICTIONS

In this section, we present two cost-sensitive re-weighting methods for imbalanced regression, which can be seamlessly integrated into the learning process of existing yield prediction models. We also provide evidence of their effectiveness in high-yield predictions.

3.1 Cost-sensitive Re-weighting Methods

The key idea is to assign a weight $w_i \in \mathcal{W}$ to each training sample $(\boldsymbol{x}_i, y_i)$, resulting in the following modified loss function $\mathcal{L}'$:

$$\mathcal{L}' = \frac{1}{N}\sum_{i=1}^{N} w_i \mathcal{L}(y_i, \hat{y}_i), \tag{1}$$

where $\mathcal{L}$ is a loss function for regression tasks, such as L_1 loss, MSE loss, and Huber loss. The distinctiveness of each re-weighting method arises from the various designs of training weights $\mathcal{W}$.

3.1.1 Focal loss. The Focal loss [7] is a specialized loss function designed to address imbalanced learning problems. It assigns varying weights to individual training samples based on their prediction difficulties. The goal is to reduce the impact of easily predictable samples while amplifying the importance of challenging samples during the training process. The weight w_i is defined as:

$$w_i = \text{SIGMOID}(\alpha \mathcal{L}(y_i, \hat{y}_i))^{\gamma} \tag{2}$$

where $\text{SIGMOID}(\cdot)$ is the sigmoid function, and α, γ are hyper-parameters. This results in the scaling factor of each training sample ranging from 0 to 1, depending on the prediction error. Notably, when $\gamma = 0$, the Focal loss is equivalent to the original loss $\mathcal{L}$. In all experiments, we set α to 0.2 and γ to 1.

Dataset	Method	MAE ↓		RMSE ↓		G-Mean ↓	
		All	Few	All	Few	All	Few
B-H	Vanilla	**4.5**±0.2	5.3±0.4	7.0±0.5	7.3±0.6	**2.5**±0.1	3.3±0.3
	+Focal	4.6±0.2	5.0±0.3	**6.5**±0.3	6.8±0.5	2.8±0.1	3.0±0.4
	+LDS	4.8±0.3	**4.6**±0.5	7.1±0.7	**6.3**±0.5	2.8±0.2	**2.8**±0.4
	+Focal+LDS	4.7±0.2	5.4±0.4	6.7±0.2	7.3±0.5	2.8±0.1	3.4±0.4
S-M	Vanilla	**7.5**±0.3	8.5±6.8	**11.1**±0.4	12.3±9.3	**4.1**±0.2	4.5±3.5
	+Focal	8.5±0.1	7.0±2.3	12.0±0.3	10.0±3.6	5.0±0.1	3.5±1.1
	+LDS	8.0±0.3	6.1±2.8	11.3±0.4	8.1±3.8	4.7±0.2	3.8±1.7
	+Focal+LDS	8.6±0.3	**5.7**±2.2	12.0±0.4	**7.8**±3.2	5.1±0.1	**3.0**±1.3
AZ	Vanilla	22.1±1.7	26.0±6.9	29.7±2.1	32.8±7.4	**12.0**±1.2	15.6±6.7
	+Focal	**22.0**±1.0	26.2±4.4	29.5±1.3	33.2±5.0	13.0±0.7	16.1±5.9
	+LDS	22.2±1.3	**24.4**±6.5	**29.1**±1.6	**30.6**±6.8	12.9±0.8	14.8±7.3
	+Focal+LDS	**22.0**±1.0	25.7±5.9	29.3±1.2	33.2±7.6	12.5±0.8	**14.2**±4.9

Table 2: Comparison of reaction yield prediction performance with and without cost-sensitive re-weighting methods on *all* and *few-shot* regions. The best results are highlighted in bold, and the second-best results are underlined.

3.1.2 Label distribution smoothing. Label distribution smoothing (LDS) [16] assigns varying weights to training samples based on their label density. The objective is to mitigate the impact of redundant samples while accentuating the significance of sparsely represented samples during training. To account for the continuity of labels, a Gaussian kernel is employed to smooth the empirical label density distribution of the label space $\mathcal{Y}$. The weight w_i for each training sample is defined as:

$$w_i = \frac{1}{\int_{\mathcal{Y}} K(y_i, y)|C_k|dy}, \tag{3}$$

where $K(y, y') = \exp\left(-\frac{\|y-y'\|^2}{2\sigma^2}\right)$ represents a Gaussian kernel with kernel size ℓ and standard deviation σ, and C_k denotes the set of training samples in the k-th bin where $y_i \in [b_{k-1}, b_k)$. In all experiments, we set ℓ to 5 and σ to 2.

3.2 Effectiveness in High-Yield Predictions

Table 2 presents the experiment results on the same three yield prediction datasets, demonstrating the effectiveness of cost-sensitive re-weighting methods (i.e., Focal loss, LDS, and the combination of both) when integrated with existing yield prediction models. Across all three datasets, imbalanced regression methods consistently achieve the best results in all 9 combinations of evaluation metrics (MAE, RMSE, and G-Mean) in the *few-shot* region. Meanwhile, the base model ("Vanilla") without any imbalanced regression designs maintains its superiority in 6 out of 9 combinations in the *all* region. This observation is understandable and aligns with the trade-off between prediction performance in underrepresented data regions and performance across the entire dataset.

However, it's important to note that while there is a performance drop in the *all* region, this drop is not significant compared to the substantial performance improvement observed in the *few-shot* region. Specifically, the average performance drops are 6.7%, 15.7%, and 0.7%, while the average performance improvements are 14.0%, 34.3%, and 7.3% on each dataset, respectively. This observation underscores the effectiveness of incorporating simple imbalanced regression methods into existing yield prediction models as a plug-in module. Furthermore, it is important to highlight that by tailoring more sophisticated imbalanced regression techniques to yield prediction, we could anticipate a further performance improvement.

4 CONCLUSION

In this paper, we have highlighted a critical issue in reaction yield prediction – the prevalent focus on achieving superior performance across the entire yield spectrum, often at the expense of overlooking yield regions with limited training samples, particularly the high-yield areas, which are of greater concern to chemists. Through extensive experiments on three real-world datasets, we have emphasized the urgent need to reframe reaction yield prediction as an imbalanced regression problem. Moreover, we have demonstrated that by incorporating simple cost-sensitive re-weighting techniques, we can significantly enhance the performance of yield prediction models in underrepresented high-yield regions.

ACKNOWLEDGMENTS

This work was supported by National Science Foundation through the NSF Center for Computer-Assisted Synthesis (C-CAS), under grant number CHE-2202693.

REFERENCES

[1] Derek T Ahneman, Jesús G Estrada, Shishi Lin, Spencer D Dreher, and Abigail G Doyle. 2018. Predicting reaction performance in C–N cross-coupling using machine learning. *Science* (2018).
[2] Paula Branco, Luís Torgo, and Rita P Ribeiro. 2017. SMOGN: a pre-processing approach for imbalanced regression. In *First international workshop on learning with imbalanced domains: Theory and applications*.
[3] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* (2002).
[4] Yu Gong, Greg Mori, and Frederick Tung. 2022. RankSim: Ranking Similarity Regularization for Deep Imbalanced Regression. In *ICML*.
[5] Zhichun Guo, Kehan Guo, Bozhao Nan, Yijun Tian, Roshni G Iyer, Yihong Ma, Olaf Wiest, Xiangliang Zhang, Wei Wang, Chuxu Zhang, et al. 2023. Graph-based molecular representation learning. In *IJCAI*.
[6] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.
[7] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *ICCV*.
[8] Gang Liu, Tong Zhao, Eric Inae, Tengfei Luo, and Meng Jiang. 2023. Semi-Supervised Graph Imbalanced Regression. In *KDD*.
[9] Michael P Maloney, Connor W Coley, Samuel Genheden, Nessa Carson, Paul Helquist, Per-Ola Norrby, and Olaf Wiest. 2023. Negative Data in Data Sets for Machine Learning Training.
[10] Damith Perera, Joseph W Tucker, Shalini Brahmbhatt, Christopher J Helal, Ashley Chong, William Farrell, Paul Richardson, and Neal W Sach. 2018. A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science* (2018).
[11] Rita P Ribeiro and Nuno Moniz. 2020. Imbalanced regression and extreme value prediction. *Machine Learning* (2020).
[12] Mandana Saebi, Bozhao Nan, John E Herr, Jessica Wahlers, Zhichun Guo, Andrzej M Zurański, Thierry Kogej, Per-Ola Norrby, Abigail G Doyle, Nitesh V Chawla, et al. 2023. On the use of real-world datasets for reaction yield prediction. *Chemical Science* (2023).
[13] Philippe Schwaller, Alain C Vaucher, Teodoro Laino, and Jean-Louis Reymond. 2021. Prediction of chemical reaction yields using deep learning. *Machine learning: science and technology* (2021).
[14] Luís Torgo, Rita P Ribeiro, Bernhard Pfahringer, and Paula Branco. 2013. Smote for regression. In *Portuguese conference on artificial intelligence*.
[15] Varvara Voinarovska, Mikhail Kabeshov, Dmytro Dudenko, Samuel Genheden, and Igor V Tetko. 2023. When yield prediction does not yield prediction: an overview of the current challenges. *Journal of Chemical Information and Modeling* (2023).
[16] Yuzhe Yang, Kaiwen Zha, Yingcong Chen, Hao Wang, and Dina Katabi. 2021. Delving into deep imbalanced regression. In *ICLR*.
[17] Dzvenymyra Yarish, Sofiya Garkot, Oleksandr O Grygorenko, Dmytro S Radchenko, Yurii S Moroz, and Oleksandr Gurbych. 2023. Advancing molecular graphs with descriptors for the prediction of chemical reaction yields. *Journal of Computational Chemistry* (2023).