Comparative Performance of EI-MS Spectrum Prediction Models under Data-scarce and Domain-imbalanced Settings

Satoki Muto, Akiko Kumada, and Masahiro Sato
Faculty of Engineering
Department of Electrical Engineering and Information Systems
The University of Tokyo
kumada@hvg.t.u-tokyo.ac.jp

Abstract

Electron ionization mass spectrometry (EI-MS) is widely used for chemical identification, yet collecting experimental spectra for newly emerging or domain-specific molecules remains time-consuming and costly. Recent chemical foundation models pre-trained on large-scale molecular corpora offer a promising approach for addressing data scarcity in such settings, but their effectiveness under domain-imbalanced conditions has not been sufficiently examined. In this study, we compare a conventional MLP-based model (NEIMS) with a fine-tuned chemical foundation model based on MolFormer-XL for EI-MS spectrum prediction under controlled few-shot conditions. Focusing on fluorine-containing molecules as a concrete example of a domain-specific subset, we vary the number of such molecules in the training data while keeping evaluation settings fixed. Across all examined conditions, the MolFormer-based model achieves higher spectral similarity and peak-level precision than NEIMS for fluorine-containing molecules. These results suggest that molecular representations learned through large-scale pre-training can be effectively leveraged for EI-MS spectrum prediction even when domain-specific training data are sparse. Our findings provide practical reference information for model selection under domain-imbalanced data conditions in automated material characterization tasks.

1 Introduction

Electron ionization mass spectrometry (EI-MS) is widely used for chemical identification in environmental analysis(Jiao et al., 2025). Recently, a variety of machine learning approaches have been proposed to predict EI-MS spectra directly from molecular structures (Wakoli et al., 2024; Overstreet et al., 2024). Among these approaches, MLP-based models have been studied as a simple and computationally efficient option for EI-MS spectrum prediction (Nguyen et al., 2024).

In parallel, recent cheminformatics research has increasingly focused on chemical foundation models pre-trained on large-scale molecular corpora (Chithrananda et al., 2020; Ahmad et al., 2022; Ross et al., 2022). These models have shown strong performance across a range of downstream chemical tasks, such as molecular property and reaction prediction (Schwaller et al., 2019), and several studies have explored their application to EI-MS spectrum prediction, including MassFormer (Young et al., 2021). However, it remains unclear how the performance of pre-trained models compares to conventional task-specific approaches under conditions in which training data are imbalanced or domain-specific molecules are sparsely observed. In practical settings, molecules belonging to specific chemical domains are often available only in limited numbers, yet systematic evaluations of model performance under such conditions remain limited. This data imbalance can substantially degrade prediction

accuracy for domain-specific molecules, highlighting the need for models that maintain performance even when such domains are sparsely represented. A notable example is fluorine-containing compounds, particularly per- and polyfluoroalkyl substances (PFAS), whose environmental distribution has become a matter of international concern. Novel PFAS compounds designed to evade regulation are often absent from existing spectral databases, and prior studies have reported degraded prediction accuracy for fluorine-containing molecules in MLP-based models (Yamaguchi et al., 2023).

In this study, we compare a conventional MLP-based model, NEIMS (Wei et al., 2019), with a fine-tuning model based on the pre-trained chemical foundation model MolFormer-XL (Ross et al., 2022) for the EI-MS spectrum prediction task under a unified evaluation setting. We adopt NEIMS as a representative fingerprint-based baseline, as it remains one of the most widely used MLP-based architectures for EI-MS prediction. To construct controlled few-shot conditions, we vary the number of fluorine-containing molecules in the training data across multiple levels, using fluorine-containing molecules as a concrete example of a domain-specific subset. We evaluate prediction performance under each condition and examine how the relative performance between the two models changes as the number of domain-specific molecules varies. Our results indicate that the MolFormer-based model achieves higher prediction performance than NEIMS across the examined settings.

2 Method and Experimental Setup

This section describes the model architectures, dataset, training protocol, and evaluation metric used in this study. We focus on comparing and analyzing the behavior of a conventional MLP-based model and a chemical foundation model under few-shot settings, where only a small number of molecules belonging to a specific chemical domain are observed in EI-MS spectrum prediction.

To ensure a fair and consistent comparison between models, we adopt a common data split and evaluation procedure across all experimental settings.

2.1 Models

As comparison targets, we consider a conventional task-specific model, NEIMS, and a fine-tuned model based on the pre-trained chemical foundation model MolFormer-XL.

NEIMS is a multilayer perceptron (MLP)-based model designed for EI-MS spectrum prediction (approximately 52M parameters). Following prior work, molecular structures are represented using extended-connectivity fingerprints (ECFP) with radius $r = 2$, with the dimensionality set to 4096 in this study.

MolFormer-XL is a chemistry foundation model pre-trained on a large-scale molecular corpus (approximately 47M pre-trained parameters, with $\sim$1.7M LoRA adapter parameters fine-tuned and a $\sim$21M prediction head trained from scratch). For the EI-MS prediction task, we fine-tune MolFormer-XL and attach a two-layer prediction head to map these embeddings to EI-MS spectra. Both models predict a 2,001-dimensional intensity vector over $m/z =$ 0–2,000 and are trained by minimizing a loss based on weighted cosine similarity.

2.2 Dataset

We use EI-MS spectra from the MassBank of North America (MoNA) database in our experiments.

After removing duplicate entries, the dataset consists of approximately 9,600 unique molecules, of which 238 (approximately 2.5%) contain fluorine atoms, illustrating the natural domain imbalance targeted in this study. The data are randomly split into 7,500 molecules for training, 500 for validation, and 1,000 for testing.

To construct few-shot settings for a specific chemical domain, we vary the number of domain-specific molecules in the training set while keeping the validation and test sets fixed. Specifically, we focus on molecules containing fluorine atoms and control the number of such

Table 1: Training hyperparameters for NEIMS and MolFormer-based models.

Hyperparameter	NEIMS	MolFormer
Optimizer	Adam	AdamW
Batch size	128	128
Learning rate	1.0×10^{-3}	1.0×10^{-3}
Weight decay	1.0×10^{-2}	1.0×10^{-2}
Dropout	0.25	0.25*
LoRA rank r / scaling α	–	16 / 32
Epochs	300	300
Early stopping patience	5	5

*Applied to prediction head only; no dropout on LoRA adapters.

molecules in the training set, denoted as X, with $X \in \{5, 10, 20, 50, 100, 175\}$. The validation and test sets contain 13 and 50 fluorine-containing molecules, respectively, and these are fixed across all conditions. Non-fluorine-containing training molecules are shared across conditions and randomly sampled, so that only the number of fluorine-containing molecules differs between settings.

2.3 Training Protocol

For each experimental condition and each run, both models are trained and evaluated using identical data splits. To account for variability due to data sampling, the training data are sampled five times for each condition, and learning and evaluation are conducted as independent runs. Final performance is reported as the average over these five runs.

The training hyperparameters for each model are summarized in Table 1. For both models, hyperparameters were selected from standard ranges commonly used in prior work and kept fixed across all experimental settings without model-specific tuning on the validation set, ensuring a fair comparison between the two models. For MolFormer, we employ parameter-efficient fine-tuning using low-rank adaptation (LoRA)(Hu et al., 2022), while keeping the remaining pre-trained parameters frozen.

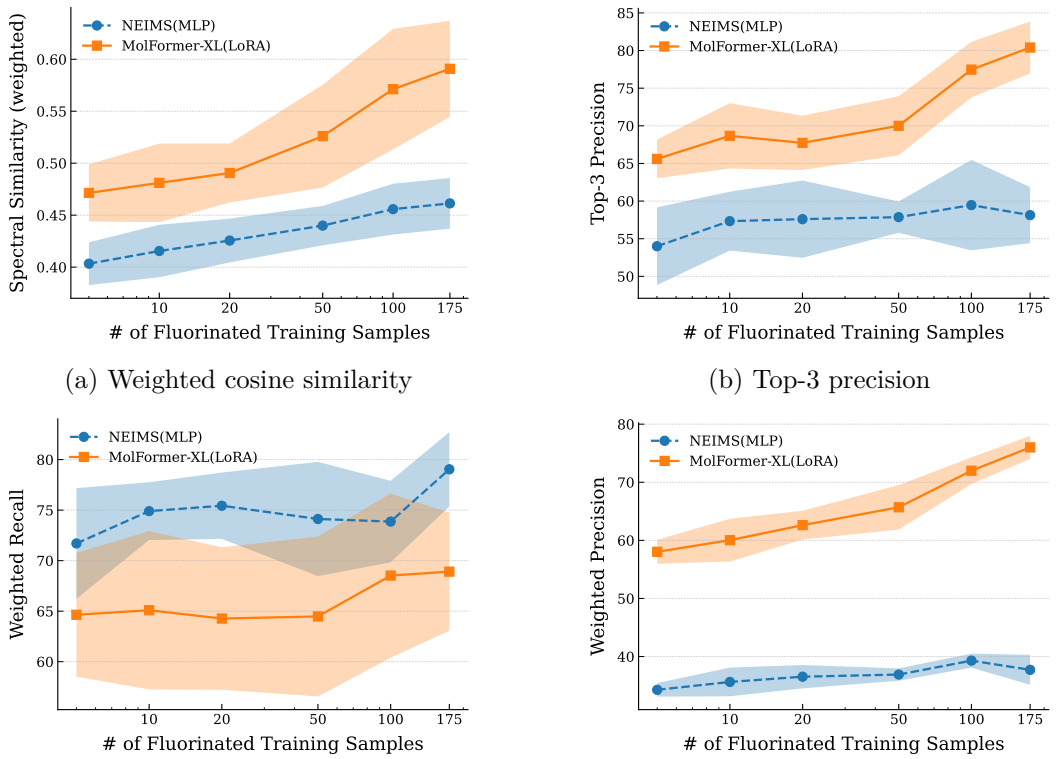
2.4 Evaluation Metrics

We evaluate the predicted EI-MS spectra using four complementary metrics. The primary metric is weighted cosine similarity, which measures spectral alignment with emphasis on high-m/z and high-intensity peaks (Young et al., 2021). We additionally report intensity-weighted precision and intensity-weighted recall, which assess peak-level prediction accuracy by comparing predicted and ground-truth peaks above an intensity threshold ($\tau = 10^{-4}$). Finally, we compute top-k precision ($k = 3$), defined as the fraction of the k highest predicted peaks that correspond to peaks present in the ground-truth spectrum.

3 Results

This section compares the EI-MS spectrum prediction performance of NEIMS and the fine-tuned model based on MolFormer-XL. Figure 1 summarizes the prediction performance for fluorine-containing molecules as a function of the number of such molecules in the training data, $X \in \{5, 10, 20, 50, 100, 175\}$. All results are reported as the mean over five independent runs.

As shown in Figure 1, the MolFormer-based model outperforms NEIMS in weighted cosine similarity, top-3 precision, and intensity-weighted precision across all values of X. In contrast, NEIMS achieves higher intensity-weighted recall, suggesting that it tends to detect a broader set of ground-truth peaks, albeit with lower precision. For both models, prediction performance generally improves as X increases.



(a) Weighted cosine similarity

(b) Top-3 precision

(c) Weighted recall

(d) Weighted precision

Figure 1: Performance metrics on fluorine-containing test molecules as a function of the number of fluorine-containing molecules in the training set ($X = 5$–175): (a) weighted cosine similarity, (b) top-3 precision, (c) intensity-weighted recall, and (d) intensity-weighted precision. Results show mean ± standard deviation over five independent runs. MolFormer-XL outperforms NEIMS in weighted cosine similarity, top-3 precision, and intensity-weighted precision across all conditions, while NEIMS achieves higher intensity-weighted recall.

Notably, the consistent performance gap across different values of X suggests that the MolFormer-based model benefits from pre-trained molecular representations across a range of domain-specific data sizes.

4 Discussion and Conclusion

In this study, we compared a conventional MLP-based model, NEIMS, with a fine-tuned model that uses molecular embeddings from the pre-trained chemical foundation model MolFormer-XL for EI-MS spectrum prediction. We examined their relative prediction performance under settings in which the number of domain-specific molecules in the training data is relatively small, focusing on fluorine-containing molecules as a concrete example.

Across the examined conditions, the MolFormer-based model achieved higher prediction performance than NEIMS in most evaluation metrics. This result indicates that molecular representations acquired through large-scale pre-training can be effectively utilized for EI-MS spectrum prediction even when only a limited number of domain-specific molecules are available for training. Since EI-MS spectra reflect molecular structure through characteristic fragmentation patterns, structural regularities learned during pre-training may contribute to improved prediction performance in such data-limited settings. In contrast, task-specific models trained from random initialization, such as NEIMS, may be more sensitive to the amount of domain-specific training data.

This study has several limitations. First, the evaluation is restricted to a single downstream task, namely EI-MS spectrum prediction, and generalization to other mass spectrometry conditions, such as different ionization methods, is not examined. Second, we focus on fluorine-containing molecules as one example of a domain-specific subset, and it remains an open question whether similar trends would be observed under alternative definitions of domain-specific molecules. Third, our comparison is limited to NEIMS as a single baseline; extending the evaluation to include more recent architectures such as MassFormer (Young et al., 2021) would further strengthen the generality of the findings.

Overall, this study provides a controlled comparison of model performance under training conditions in which domain-specific molecules are limited in number for EI-MS spectrum prediction, and suggests that leveraging large-scale pre-training on readily available molecular data can serve as a practical strategy for mitigating data scarcity in experimental materials science tasks. The observations reported here offer basic reference information for considering model choice under practical data conditions involving domain imbalance.

Reproducibility

Code is available at https://github.com/0tum/eiMS-fewshot-benchmark.

References

Walid Ahmad, Elana Simon, Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta-2: Towards chemical foundation models. arXiv preprint arXiv:2209.01712, 2022.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. arXiv preprint arXiv:2010.09885, 2020.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. ICLR, 1(2):3, 2022.

Zhaoyu Jiao, Sachi Taniyasu, Nanyang Yu, Xuebing Wang, Nobuyoshi Yamashita, and Si Wei. Two-layer homolog network approach for pfas nontarget screening and retrospective data mining. Nature Communications, 16(1):688, 2025.

Julia Nguyen, Richard Overstreet, Ethan King, and Danielle Ciesielski. Advancing the prediction of ms/ms spectra using machine learning. Journal of the American Society for Mass Spectrometry, 35(10):2256–2266, 2024.

Richard Overstreet, Ethan King, Grady Clopton, Julia Nguyen, and Danielle Ciesielski. Qc-gn2oms2: a graph neural net for high resolution mass spectra prediction. Journal of Chemical Information and Modeling, 64(15):5806–5816, 2024.

Jerret Ross, Brian Belgodere, Vijil Chenthamarakshan, Inkit Padhi, Youssef Mroueh, and Payel Das. Large-scale chemical language representations capture molecular structure and properties. Nature Machine Intelligence, 4(12):1256–1264, 2022.

Philippe Schwaller, Teodoro Laino, Théophile Gaudin, Peter Bolgar, Christopher A Hunter, Costas Bekas, and Alpha A Lee. Molecular transformer: a model for uncertainty-calibrated chemical reaction prediction. ACS central science, 5(9):1572–1583, 2019.

Julia Wakoli, Afia Anjum, Tanvir Sajed, Eponine Oler, Fei Wang, Vasuk Gautam, Marcia LeVatte, and David S Wishart. Gcms-id: a webserver for identifying compounds from gas chromatography mass spectrometry experiments. Nucleic Acids Research, 52(W1): W381–W389, 2024.

Jennifer N Wei, David Belanger, Ryan P Adams, and D Sculley. Rapid prediction of electron–ionization mass spectrometry using neural networks. ACS central science, 5 (4):700–708, 2019.

Ryohei Yamaguchi, Shigenori Takeda, Masatsugu Yamada, Toshifumi Kakiuchi, and Yutaka Imamura. Improving deep neural network in predicting electron ionization mass spectra with molecular similarity-wise sampling. AGC research report, 73:42–49, 2023.

Adamo Young, Bo Wang, and Hannes Röst. Massformer: Tandem mass spectrum prediction for small molecules using graph transformers. arXiv preprint arXiv:2111.04824, 2021.