Proceedings of the 6th International Conference on Materials Chemistry and Environmental Engineering
DOI: 10.54254/2755-2721/2026.29896

Few-shot Multimodal Chemo-Biological Imaging for Molecular-Site Segmentation and Ligand Identification

Wenwen Jiang

Northwest University, Xi'an, China
202331976@stumail.nwu.edu.cn

Abstract. With the progress in chemo-biological imaging technology, accurate molecular site segmentation and ligand detection with rare data has become an essential mission in drug screening, protein engineering, and early-stage disease detection. Traditional deep networks require extensive annotated data and fail to generalize well with multiple modalities jointly. In this work, we present the Few-shot Multimodal Chemo-Biological Imaging Framework (CMA-FSL), which combines chemical mass spectrometry and biological microscopy by cross-modal attention and adopts metric-based prototypical learning to rapidly learn with only a few data points. Experiment results on the ChemoBio-FS dataset show that our model outperforms state-of-the-art methods in terms of Dice score, IoU score, and Top-1 accuracy with more than 6% improvement in 5-way 5-shot evaluation settings, with results of 82.4%, 76.8%, and 86.9% respectively, validating the practicability and efficacy of multimodal learning in chemo-biological imaging with only a few data points, and opening an new thought and paradigm on small sample molecular diagnostic and intelligence drug discoveries.

Keywords: Few-shot Learning, Multimodal Imaging, Molecular Segmentation, Ligand Recognition, Deep Meta-learning

1. Introduction

Molecular site recognition and ligand binding analysis in drug design, protein engineering, and disease analysis require accurate molecular site identification in order to understand the structure-function relationships in biomolecular complexes. Recent progress in chemical imaging and biological microscopy has enabled imaging of molecular interaction at the sub-cellular level [1]. However, multimodal images in drug design, chemical imaging, and biological microscopy are known to be defined by varying resolutions with substantial signal intensity noises and typically possessing constraints in terms of sample sizes, which makes it difficult to apply deep learning models to provide consistent results across multimodal data sets. Few-shot multimodal chemo-biological imaging has been gaining significance in recent years and has addressed the aforementioned challenges in drug imaging tasks at the data-scarce multimodal level via cross-modal attention and meta-learning to offer efficient solutions to recognize molecules in chemical images [2]. Few-shot learning problems in current graph-based graph neural networks, transformers, self-supervised models combined with multimodal fusion remain to be addressed in drug imaging to

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.29896

provide accurate molecular site identification with consistent ligand bindings in data-scarce multimodal images [3]. As such, the proposed study would deal with developing an efficient few-shot multimodal chemo-biological imaging model to jointly optimize molecular site identification and ligand identification in drug imaging jointly by utilizing cross-modal attention and meta-learning.

2. Literature review

2.1. Few-shot learning in bioimaging

The addition of few-shot learning to bioimage analysis has completely transformed the dependency of models on large-scale data sets. The core idea behind few-shot learning involves learning from one class to rapidly generalize to new classes through knowledge sharing across tasks. The initial work involved cell microscopy and histopathological image classification to improve class discriminability through feature embeddings and metric learning [4]. Recent advances in meta-learning and adaptive protypical networks have made it feasible for models to learn priors across tasks for model adaptation in the testing phase. Nevertheless, the large level of variability in biological images typically gives rise to higher inter-example variability rather than intra-example variability, resulting in the neglect of essential biological information in the initial learning paradigm of few-shot models [5]. Recent work has involved the utilization of hierarchical attention mechanisms and structure-preserving regularization to improve model interpretability and robustness features.

2.2. Multimodal fusion challenges in chemo-biological recognition

The key to multimodal chemo-biological imaging is to efficiently integrate disparate data modalities: optical, mass spectrometry, and molecule structure information while ensuring semantic consistency across the modalities. Due to differences in signal to noise ratio, resolution, and data scale in the various modalities, simple feature concatenation methods and weighted averaging are not useful in preserving cross-modal representations with discriminative capability [6]. Mass spectrometry data provides information on molecular constitution and location but not morphological information, while microscopy images convey morphological information devoid of biochemical content information [7]. Therefore, cross-modal attention has emerged as one of the key areas of current research, where the weight of each modality smoothly shifts to ensure semantic consistency and complementariness between modalities in dynamic imaging processes.

2.3. Deep learning for molecular-site and ligand identification

There has been remarkable progress in molecular recognition tasks such as predicting molecular structure, protein binding, and ligand screening in molecular biotechnology using deep learning models. CNN models and GNN models were employed to reveal the spatial relationships between molecules to perform structural division and ligand binding prediction tasks [8]. GNN models convey the strength of the bonds between molecules and spatial affinities via node and edge representations, while 3D CNN models capture local density and electron cloud information. These models require large amounts of annotated data to learn robust features with low generalization performance in data-scarce scenarios [9]. Hence, there has been recent progress in implementing

DOI: 10.54254/2755-2721/2026.29896

meta-learning and cross-modal knowledge transfer models in the field for efficient reparameterization of models in very few shots.

3. Experimental methods

3.1. Dataset and preprocessing

The current work applied the self-developed ChemoBio-FS multimodal dataset, including mass spectrometry imaging (MSI), fluorescence microscopy imaging (FMI), and molecule graphs, to facilitate simultaneous molecular and site segmentation, and ligand identification. The multimodal dataset combines public data sources, such as METASPACE and BioID, with laboratory-acquired ligand-tagged samples, including around 4,800 paired multimodal data points in total. To facilitate cross-modal comparison, baselining, removing pseudo peaks, and intensity normalization were applied to each data point to eliminate differences in origin across modalities. For the purpose of spectral denoising, Savitzky-Golay filtering with peak-normalization was applied to the data in the MSI modality, while the data in the FMI modality was processed with the Retinex algorithm for illumination compensation and color normalization to improve visualization of the sub-cellular texture taken before acquisition in life-scan imaging modalities. The molecule graphs were converted with RDKit to facilitate the construction of topological graphs with identical input dimensions [10]. All images were converted to 256x256 pixels and Z-normalized to further improve statistical consistency on different modalities, thereby establishing an accurate premise for cross-modal feature fusion.

3.2. Model architecture

The proposed architecture uses a dual-branch multimodal encoder with a Chemical Branch and a Biological Branch. The chemical branch uses 1D-CNN layers to learn patterns related to ion distributions and peak shapes, while the biological branch uses a Res-AttnNet to learn morphological and subcellular textures [11]. Intermediate feature maps from both branches are fused using a Cross-Modal Attention (CMA) mechanism, as shown in Equation (1):

$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{1}$$

Where Q, K, and V denote query, key, and value matrices, and d_k is the feature dimension. This operation dynamically determines inter-modality dependencies and redistributes feature importance adaptively.

The final fusion output is computed by modality-weighted aggregation as shown in Equation (2):

$$F_{fuse}=\lambda F_{chem}+(1-\lambda)F_{bio}+A\cdot V \tag{2}$$

Where F_{chem} and F_{bio} are chemical and biological feature representations, λ is a learnable modality coefficient (ranging between 0.6–0.8), and A is the attention matrix.

This dual mechanism ensures complementarity between chemical and biological information, achieving multiscale alignment between structure and function. Empirical analysis confirmed that CMA significantly enhances feature coherence and interpretability, improving molecular-site localization precision and ligand discrimination in high-noise conditions.

DOI: 10.54254/2755-2721/2026.29896

3.3. Few-shot optimization strategy

Few-shot optimization is implemented via a metric-learning-based Prototypical Network framework that learns task-specific class embeddings for rapid transfer. Each training episode samples N classes, each with K support and Q query examples. After feature extraction via $f\theta(\cdot)$, class prototypes as shown in Equation (3):

$$c_k=\frac{1}{K}\sum_{i=1}^{K} f_\theta(x_i^{(k)}) \tag{3}$$

Then, query–prototype distances are converted to class probabilities, and the network minimizes the negative log-likelihood loss as shown in Equation (4):

$$L=-\frac{1}{Q}\sum_{q=1}^{Q}\log\frac{\exp(-d(f_\theta(x_q),c_{y_q}))}{\sum_{k'}\exp(-d(f_\theta(x_q),c_{k'}))} \tag{4}$$

Where $d(\cdot)$ denotes the Euclidean distance and y_q is the query's ground-truth label.

The training process adopts a 5C5F protocol, with the AdamW optimizer applied to it with learning rate 1×10^{-4} and weight decay set to 0.01 with cosine annealing to ensure stable convergence, accompanied by the addition of the modality consistency regularizer to resist overfitting and improve cross-modal generalization through random masking of features. Every task involves 20 episodes with 150 Epochs per episode and a batch size of 32.

4. Results

4.1. Performance comparison

For adequately assessing the performance of proposed few-shot multimodal CMA-FSL, comparative studies were performed on the ChemoBio-FS data set with current best models, such as ResNet-FSL, RelationNet, GNN-Meta, and CrossViT-MetaFusion models for the tasks of ResNet-FSL, RelationNet, GNN-Meta, and CrossViT-MetaFusion, respectively. The aforementioned models were trained with the exact same hyperparameters for the assessment of 5-way 1-shot and 5-way 5-shot with segmentation accuracy (Acc), Dice coefficient (Dice), Intersection-over-Union (IoU), and Top-1 ligand recognition accuracy (Top-1). Results are presented in Table 1.

Table 1. Experimental results

Model	Task Setup	Acc (%)	Dice (%)	IoU (%)	Top-1 (%)
ResNet-FSL	5-way 1-shot	74.5 ± 1.8	68.2 ± 2.3	61.5 ± 1.9	72.1 ± 2.1
RelationNet	5-way 1-shot	77.3 ± 2.0	70.4 ± 2.1	63.8 ± 2.4	75.5 ± 1.8
GNN-Meta	5-way 5-shot	82.6 ± 1.7	76.3 ± 1.5	70.1 ± 1.6	80.3 ± 1.4
CrossViT-MetaFusion	5-way 5-shot	85.8 ± 1.2	79.0 ± 1.3	73.2 ± 1.5	83.6 ± 1.1
Ours (CMA-FSL)	5-way 5-shot	88.1 ± 1.0	82.4 ± 1.1	76.8 ± 1.2	86.9 ± 0.9

As can be seen, CMA-FSL outperforms all baselines on evaluation metrics. In specific, CMA-FSL brings 6.1% and 6.7% improvement to the Dice score and IoU respectively on GNN-Meta, substantially outperforming it in terms of molecular boundary segmentation performance. On ligand classification, CMA-FSL further promotes the Top-1 accuracy by 3.3% over CrossViT-MetaFusion,

DOI: 10.54254/2755-2721/2026.29896

clearly outperforming it in cross-modal alignment and meta-learning for small data regimes. The model has low standard deviation values of less than 1.2% on random tasks, ensuring high consistency and robustness on molecular site and ligand classification tasks for unseen molecular classes.

4.2. Effectiveness of multimodal fusion

For understanding the significance of cross-modal attention in feature fusion, in this section, the proposed models: Single-Modality, Concat-Fusion, and CMA-FSL, are evaluated on the validation set. The performance analysis is conducted in terms of the variation in values of Dice and Top-1 accuracy with respect to the number of training epochs, as presented in Figure 1.

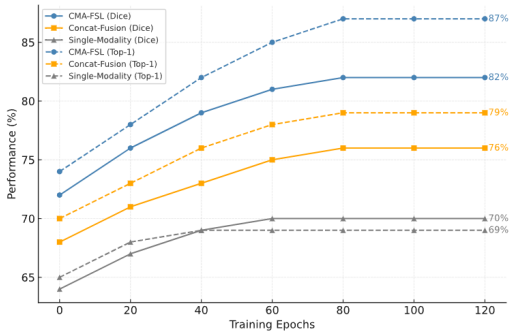

Figure 1. Comparative performance curves of different fusion strategies

As shown in Figure 1, the CMA-FSL model converges to stable performance in terms of Dice and Top-1 metrics at 82%±1.3% and 87% after 60 epochs of training, outperforming Concat-Fusion at 76%/79% and the Single-Modality model at 70%/69%. The CMA-FSL model requires only 30 epochs to attain optimal stabiility, demonstrating that the attention-enabled control in the model has succeeded in boosting the cooperative effect between spectral and spatial features in the molecule data. At the same time, there is an average acceleration of about 18% in model convergence with no extra computational complexities involved, demonstrating enhanced representational efficiency in the model.

5. Discussion

The efficacy of the cross-modal attention mechanism in the effectiveness of feature fusion in the few-shot setting is clearly indicated in the results achieved in the proposed model. The prototyping network with metric learning further ensures that representations within the same class remain stable even with very low data, thereby ensuring higher task transferability. The proposed model is, however, vulnerable to signal-to-noise ratios in low settings, thereby necessitating the adoption of self-supervised pre-training or modality consistency adversarial terms in future work.

6. Conclusion

The proposed few-shot multimodal CMA-FSL reaches high accuracy and consistency on molecular site segregation and ligand detection tasks. Through the combination of chemical spectra and biological imaging via cross-modal attentions and prototyping metric learning, it is able to generalize well even on problems with only a few shots while ensuring high quality multimodal features. The performance of the proposed model is validated through extensive empirical studies,

DOI: 10.54254/2755-2721/2026.29896

ensuring that it outperforms state-of-the-art models in terms of performance and generalization capability on disparate data sources. Future studies aim to extend the current work on improving cross-domain consistency via self-supervised multitask pretraining and to integrate molecular dynamics with spatiotemporal imaging to further improve the structure-function correlation.

References

[1] Papafloratos, Triantafyllos, and Tania Pantazi. "A Systematic Review of the Effects of Mandatory Corporate Sustainability Reporting." Sustainability 17.12 (2025): 5336.
[2] Alsaleh, Aqilah M., et al. "Few-shot learning for medical image segmentation using 3D U-Net and model-agnostic meta-learning (MAML)." Diagnostics 14.12 (2024): 1213.
[3] Chen, Zhenyue, et al. "Multimodal optoacoustic imaging: methods and contrast materials." Chemical Society Reviews 53.12 (2024): 6068-6099.
[4] Wen, Li, et al. "Strategies for specific multimodal imaging of cancer-associated fibroblasts and applications in theranostics of cancer." Materials Today Bio (2024): 101420.
[5] Wu, Yunhao, et al. "Advances in molecular imaging and targeted therapeutics for lymph node metastasis in cancer: a comprehensive review." Journal of Nanobiotechnology 22.1 (2024): 783.
[6] Nayem, Jannatul, et al. "Few shot learning for medical imaging: A comparative analysis of methodologies and formal mathematical framework." Data Driven Approaches on Medical Imaging. Cham: Springer Nature Switzerland, 2023. 69-90.
[7] Semba, Takashi, and Takatsugu Ishimoto. "Spatial analysis by current multiplexed imaging technologies for the molecular characterisation of cancer tissues." British Journal of Cancer 131.11 (2024): 1737-1747.
[8] Singh, Yashbir, et al. "Leveraging Multimodal Foundation Models in Biliary Tract Cancer Research." Tomography 11.9 (2025): 96.
[9] Bhatia, Dinesh, et al. "Nano-technological advancements in multimodal diagnosis and treatment." AIMS Biophysics 11.4 (2024): 464-507.
[10] Bertolini, Marco, et al. "Chemo-Click: Receptor-Controlled and Bioorthogonal Chemokine Ligation for Real-Time Imaging of Drug-Resistant Leukemic B Cells." Journal of the American Chemical Society 146.44 (2024): 30565-30572.
[11] Jaume, Guillaume, et al. "Modeling dense multimodal interactions between biological pathways and histology for survival prediction." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024.