iMML: A Python package for multi-modal learning with incomplete data

Alberto López[1*,2] **John Zobolas**[1,2] **Tanguy Dumontier**[1] **Tero Aittokallio**[1,2,3,4]
[1]Oslo Centre for Biostatistics and Epidemiology (OCBE), University of Oslo, Norway
[2]Department of Cancer Genetics, Institute for Cancer Research, Oslo University Hospital, Norway
[3]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Finland
[4]iCAN Digital Precision Cancer Medicine Flagship, Helsinki University Hospital, Finland
{a.l.sanchez,ioannis.zompolas,t.a.aittokallio}@medisin.uio.no

Abstract

Machine learning models that effectively integrate multiple data modalities generally outperform their uni-modal counterparts. However, in many situations, certain modalities or variables are missing for a subset of samples, leading to a limited performance or failure of conventional methods. This has given a rise to the field of multi-modal learning with incomplete data, an area that has grown rapidly due to its broad real-world applications. Despite this, the community still lacks standardized tools to effectively handle incomplete multi-modal data. To fill this gap, we developed iMML, a unified, user-friendly Python package with methods designed for integrating, processing, and analyzing incomplete multi-modal data. iMML is available at `https://github.com/ocbe-uio/imml`, with fully open-source (BSD-3 license), and a user documentation at `https://imml.readthedocs.io/`.

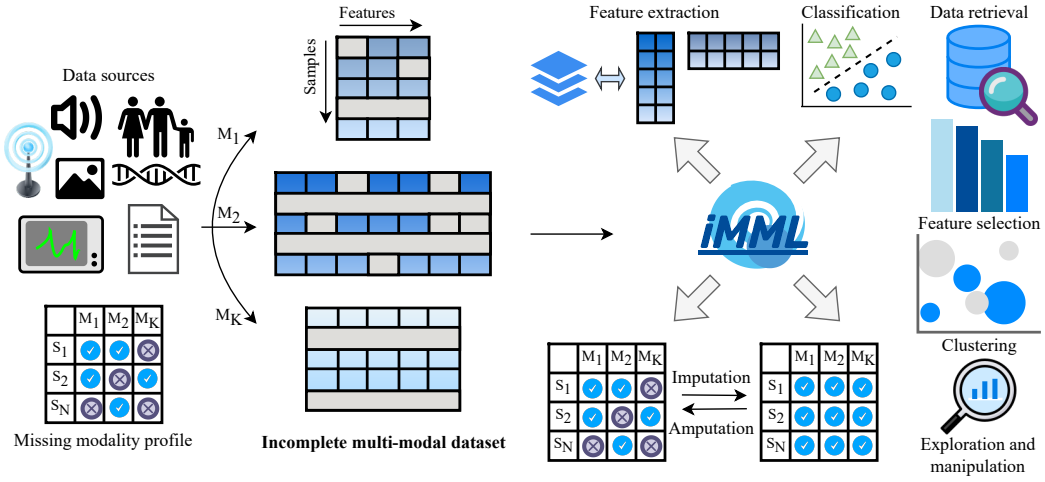
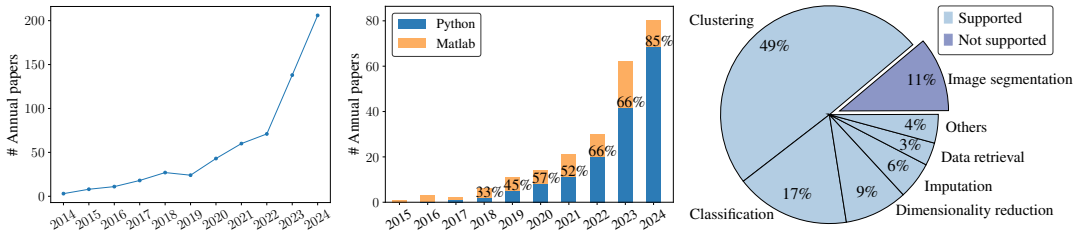


Figure 1: **Overview of iMML.** iMML provides a robust tool-set for analyzing incomplete multi-modal datasets to support a wide range of real-world machine learning tasks.

1 Introduction

Multi-modal learning (MML), where diverse data types are integrated and analyzed together, has emerged as a critical field in artificial intelligence. However, multi-modal datasets are often incomplete

39th Conference on Neural Information Processing Systems (NeurIPS 2025).



(a) Annual number of publications. (b) Publications with code. (c) Tasks.

Figure 2: **Growth in the field of multi-modal learning with incomplete data**. (a) Number of publications is increasing almost exponentially. (b) Matlab dominated the early years, but Python has since surpassed it. (c) iMML supports the most common tasks.

due to various reasons, such as measurement failures, hardware restrictions, privacy limitations or high costs of data collection [1]. For instance, The Cancer Genome Atlas (TCGA, a widely-used resource for cancer research) contains numerous missing omics for a subset of the patients [2].

Learning from incomplete multi-modal datasets (IMMD) has seen a significant growth over the last years (Figure 2) [3]. Numerous recent papers have identified the field as a promising area for future research [4, 5, 6, 7, 8, 9]. Adapting existing approaches to handle IMMD has also been widely suggested as a future work [10, 11, 12]. Despite this progress, several limitations still persist. The landscape of available methods is fragmented, largely due to the diversity of application cases and data modalities. The systematic application and comparison of the current methods are often hindered by practical challenges, such as incompatible input data formats, outdated requirements and conflicting software dependencies. As a result, researchers frequently face challenges in choosing a practical method and invest considerable efforts into reconciling codebases, rather than addressing the core scientific questions. Moreover, most methods reported in the literature are rarely used in applied settings, despite the ubiquity of IMMD. This suggests that the community currently lacks robust and standardized tools to effectively handle IMMD.

To address this gap, we have developed iMML (compared with other popular open-source packages in Table 1), an open-source Python package designed for multi-modal learning with incomplete data. The key features of this package are:

- **Coverage**: More than 25 methods available for integrating, processing, and analyzing incomplete multi-modal datasets, implemented as a single, user-friendly interface.
- **Comprehensive**: Designed to be compatible with widely-used machine learning and data analysis tools, allowing use with minimal programming effort. An extensive documentation enables end-users to apply its functionality effectively.
- **Extensible**: A unified framework where researchers can contribute and integrate new approaches, serving as a community platform for hosting new methods.

2 The iMML library

iMML is organized into several modules, each designed to address a specific task: **Ampute**: novel functions for testing MML methods by simulating missing modalities based on four missingness patterns; **Classify**: several deep learning models for classification tasks, based on transformer architectures that leverage pretrained models; **Cluster**: multiple methods that utilize various approaches, including deep learning, matrix factorization, kernel methods and graph learning for clustering; **Decomposition**: this module facilitates feature extraction by transforming the original feature space into a more compact representation; **Explore**: a diverse set of tools to explore a multi-modal dataset; **Load**: classes to load datasets for deep learning algorithms; **Impute**: designed for filling missing data, which can be particularly useful when using external methods that are unable to handle missing values directly; **Feature selection**: this module enables the identification of key features; **Preprocessing**: classes for processing datasets for downstream tasks; **Retrieve**: tools for extracting information from storage systems; **Statistics**: functions for computing multi-modal data statistics, such as the

Table 1: **Use cases supported by iMML and other open source libraries** (if not otherwise specified in the table, the package is implemented in Python). Abbreviations: UM, Uni-modal; MM: Multi-modal; IMM: Incomplete multi-modal.

PACKAGE	IMPUTATION		AMPUTATION		MACHINE LEARNING		
	UM	MM	UM	MM	UM	MM	IMM
iMML		✓		✓		✓	✓
[13] (MATLAB)						✓(CLUSTER)	✓(CLUSTER)
MULTIZOO (2023)						✓	
TORCHMULTIMODAL (2022)						✓	
SCIKIT-MULTIMODALLEARN (2022)						✓	
HYPERIMPUTE (2022)	✓						
PYAMPUTE (2022)			✓				
MVLEARN (2021)						✓	
MDATAGEN (2019)	✓						
SCIKIT-LEARN (2011)	✓				✓		

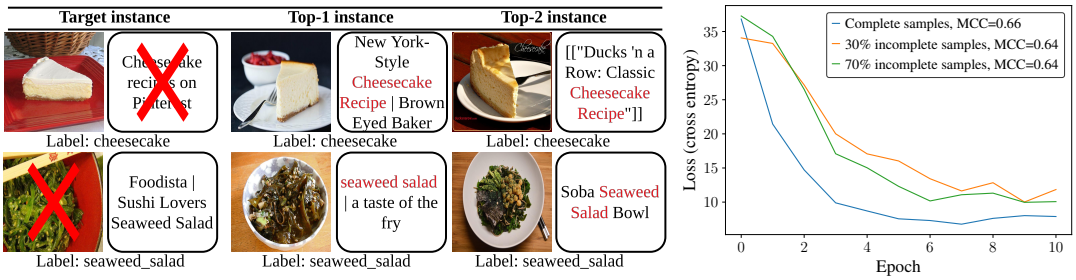
redundancy, uniqueness and synergy of the modalities; **Utils**: utilities for data manipulation, such as input validation; **Visualize**: functions to visually explore a multi-modal dataset.

3 Illustrative case studies

We present several use cases to showcase the versatility of iMML. In the first case, we used a reduced version of the Food101 dataset, an image-text dataset [14], for robust retrieval and classification. We first built a multi-channel retriever, which effectively identified relevant cross-modal instances (Figure 3a). Subsequently, we trained a classification model using RAGPT [15], which demonstrated strong robustness and similar performance (using Matthews correlation coefficient, MCC) on the test set under complete data, 30% and 70% missing modalities (Figure 3b).

For the second case, we showed how iMML simplifies performance evaluation by simulating block-wise IMMD. This so-called data amputation process allows for controlled testing of methods by generating missing data from various missingness patterns, thereby reflecting real-world scenarios, where different data modalities may be either partially observed or entirely missing.

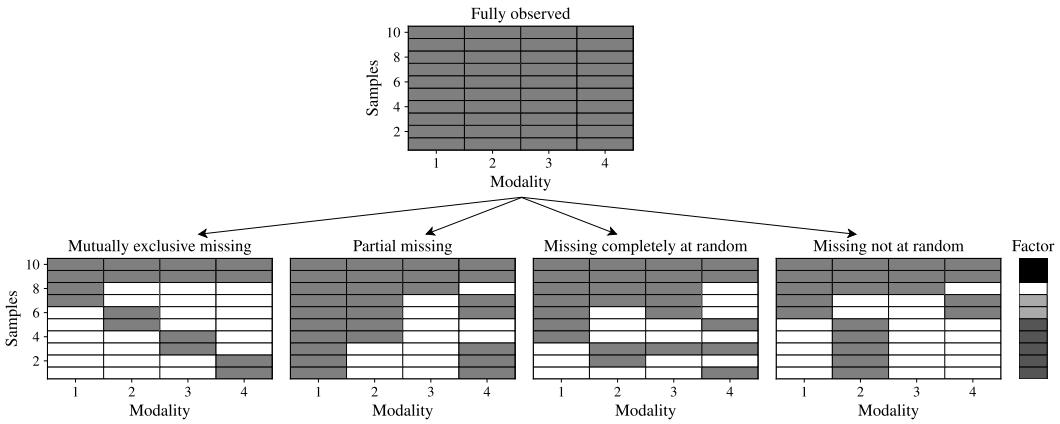
In the last case, we focused on dimensionality reduction, a typical task in biomedical research where high-dimensional datasets are particularly common. We used the nutrimouse dataset [16], which includes gene expression levels and concentrations of fatty acids measured in 40 mice, each labeled by the genetic type. We first simulated block- and feature-wise missing data to reflect real-world biomedical datasets. We then applied the jNMF algorithm [17] for feature extraction and feature selection, followed by a genetic classification task. For comparison, we also included baselines using randomly selected features and all available features. The extracted features demonstrated strong robustness (Figures 5a), while the selected features achieved impressive accuracy, particularly



(a) Retrieved instances.

(b) Classification performance.

Figure 3: **Classification and retrieval on an incomplete vision-language dataset**. (a) Top-2 retrieved instances when using image-only (top) and text-only (bottom) as target instances. (b) Test set loss during training, with performance (MCC) on the test set.



(a) Visual representation of the amputation patterns.

PATTERN	# COMPLETE SAMPLES	# INCOMPLETE SAMPLES	# SAMPLES PER MODALITY
MEM	2	8	[4, 4, 4, 4]
PM	2	8	[10, 7, 3, 7]
MCAR	2	8	[7, 6, 6, 5]
MNAR	2	8	[5, 8, 3, 4]

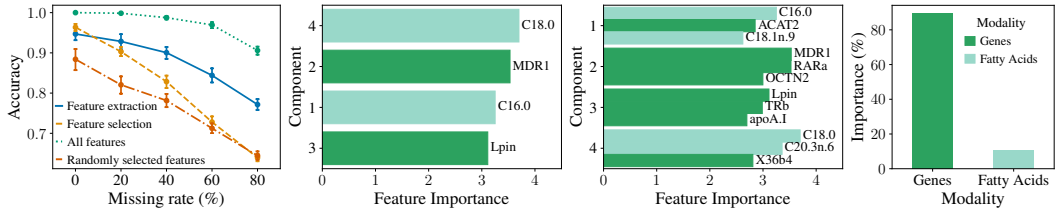
(b) Report of amputation patterns.

Figure 4: **Block-wise missing data generation**. (a) Visual representation of the misingness patterns (white areas represent missing data). (b) A summary of the number of complete and incomplete samples in each scenario, along with the number of observed samples in each modality.

when the missing rates were not very high (<40%). The top features were originated from both gene expression and fatty acid measurements, with two genes and two fatty acid identified as the most important features, hence forming a multi-modal biomarker panel (Figure 5b). The extracted features, representing aggregated versions of the original features, were further decomposed to gain insights into the model's behavior (Figure 5c). Quantifying the relative importance of the modalities revealed that gene expression contributed much more than fatty acid measurements. (Figure 5d)

4 Conclusions

iMML is the most comprehensive open-source library for multi-modal learning with incomplete data. Our vision is to establish iMML as a leading library for MML. We welcome the open-source community to contribute and help us extend the library with the emerging methods. Such a community-wide effort will make iMML more powerful for the machine learning community.



(a) Performance. (b) Top features. (c) Explained components. (d) Importance.

Figure 5: **Feature extraction and selection on an incomplete multi-modal biomedical dataset**. (a) Classification accuracy using extracted features, selected features, all features and randomly selected features across varying missing rates (0% to 80%). (b) Top features selected when the missing rate was 20%. (c) Decomposition of the extracted features showing their most influential features at a 20% missing rate. (d) Modality importance with a 20% missing rate.

References

[1] Renjie Wu, Hu Wang, and Hsiang-Ting Chen. A comprehensive survey on deep multimodal learning with missing modality, 2024.

[2] Galadriel Briere, Elodie Darbo, Patricia Thebault, and Raluca Uricaru. Consensus clustering applied to multi-omics disease subtyping. *BMC Bioinformatics*, 22, 07 2021.

[3] Jingjing Tang, Qingqing Yi, Saiji Fu, and Yingjie Tian. Incomplete multi-view learning: Review, analysis, and prospects. *Applied Soft Computing*, 153:111278, 03 2024.

[4] Jing Zhao, Xie Xijiong, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38, 02 2017.

[5] Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu. Deep multi-view learning methods: A review. *Neurocomputing*, 448:106–129, 2021.

[6] Adrienne Kline, Hanyin Wang, Yikuan Li, Saya Dennis, Meghan Hutch, Zhenxing Xu, Fei Wang, Feixiong Cheng, and Yuan Luo. Multimodal machine learning in precision health: A scoping review. *npj Digital Medicine*, 5(1):171, 2022.

[7] M. Khojaste-Sarakhsi, Seyedhamidreza Shahabi Haghighi, S.M.T. Fatemi Ghomi, and Elena Marchiori. Deep learning for alzheimer's disease diagnosis: A survey. *Artificial Intelligence in Medicine*, 130:102332, 2022.

[8] Sandra Steyaert, Marija Pizurica, Divya Nagaraj, Priya Khandelwal, Tina Hernandez-Boussard, Andrew J Gentles, and Olivier Gevaert. Multimodal data fusion for cancer biomarker discovery with deep learning. *Nature machine intelligence*, 5(4):351–362, 2023.

[9] Uno Fang, Man Li, Jianxin Li, Longxiang Gao, Tao Jia, and Yanchun Zhang. A comprehensive survey on multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12350–12368, 2023.

[10] Dong Huang, Chang-Dong Wang, and Jian-Huang Lai. Fast multi-view clustering via ensembles: Towards scalability, superiority, and simplicity. *IEEE Transactions on Knowledge and Data Engineering*, 35(11):11388–11402, 2023.

[11] Man-Sheng Chen, Chang-Dong Wang, and Jian-Huang Lai. Low-rank tensor based proximity learning for multi-view clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(5):5076–5090, 2023.

[12] Chang Tang, Zhenglai Li, Jun Wang, Xinwang Liu, Wei Zhang, and En Zhu. Unified one-step multi-view spectral clustering. *IEEE Transactions on Knowledge and Data Engineering*, 35(6):6449–6460, 2023.

[13] Jie Wen, Zheng Zhang, Lunke Fei, Bob Zhang, Yong Xu, Zhao Zhang, and Jinxing Li. A survey on incomplete multiview clustering. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 53(2):1136–1149, 2023.

[14] Xin Wang, Devinder Kumar, Nicolas Thome, Matthieu Cord, and Frederic Precioso. Recipe recognition with large multimodal food dataset. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–6. IEEE, 2015.

[15] Jian Lang, Zhangtao Cheng, Ting Zhong, and Fan Zhou. Retrieval-augmented dynamic prompt tuning for incomplete multimodal learning. *arXiv preprint arXiv:2501.01120*, 2025.

[16] Pascal G. P. Martin, Hervé Guillou, Frédéric Lasserre, Sébastien Déjean, Annaïg Lan, Jean-Marc Pascussi, Magali Sancristobal, Philippe Legrand, Philippe C. Besse, and Thierry Pineau. Novel aspects of pparα-mediated regulation of lipid and xenobiotic metabolism revealed through a nutrigenomic study. *Hepatology*, 45, 2007.

[17] Koki Tsuyuzaki and Itoshi Nikaido. nntensor: An r package for non-negative matrix/tensor decomposition. *Journal of Open Source Software*, 8(84):5015, 2023.