**Abstract Number :** abs25-12042

**Abstract Title :** Marvin: Deep generative model with structured latent space for flow cytometry cell classification and discovery of new population

**Category:** 800s - Gene Therapies, Chemical Biology, and Emerging Diagnostics

**Review Category:** Emerging Tools, Techniques, and Artificial Intelligence in Hematology

**Authors**

Adrien De Voeght[1, 2], Fanny Bodart[1], Frederic Baron[2], Gilles Louppe[1]

[1] University Of Liège, Montefiore Institute, Department of Electrical Engineering and Computer Science, Artificial Intelligence and Deep Learning, Liège, Belgium, [2] University Of Liège, Centre Hospitalier Universitaire de Liège, Clinical Hematology, Liège, Belgium

**Abstract Body**

BACKGROUND

Mass or flow cytometry are used in research and clinical settings to analyze cell populations. It generates multidimensional data representing many cells with multiple markers. Cytometry data analysis relies on manual gating strategies that are time-consuming and operator-dependent. Several machine learning methods have been developed to automate cell classification, but most focus solely on classification of known populations. Generative deep learning (DL) models offer potential advantages for simultaneously handling both classification and discovery tasks in cytometry data.

AIMS

We aim to achieve three complementary goals through generative DL: accurate classification of known immune cell populations, discovery of novel cell subpopulations and identification of interactions between immune cell populations across different experimental conditions.

METHODS

**Model**: We develop MARVIN: a Mixture-based Variational Autoencoder designed to model cytometry data through a structured latent representation conditioned on cell type. Each data point (cell markers) is assigned to a specific latent Gaussian distribution, allowing the model to associate to each cell population its own latent component. This enables simultaneously the classification of known cell populations and the discovery of novel subpopulations within a unified framework.
Our model operates in a semi-supervised framework, meaning that it is inherently capable of handling missing labels (cells left unannotated by experts). It can be used in scenarios where only a single patient's cell populations have been gated, while the remaining patients are entirely unlabeled, leading to a time-saving benefit for the practitioner.

**Datasets**: We evaluate our approach on publicly available datasets: (1) POISED (30 samples, 15 patients, peanut-stimulated vs. unstimulated conditions, 22 immune cell classes, 4,178,320 cells); (2) AML (2 donor samples, 14 immune populations, 104,184 cells); and (3) BMMC dataset (13 markers, 19 cell types, 61,725 cells) and 1 in-house MRD dataset (6,667,147 cells, of which 10,222 are leukemic cells).

**Evaluation**: We assess model performance using accuracy, balanced accuracy, and F1 score, with emphasis on the latter two metrics due to the unbalanced distribution of cell populations. We compare MARVIN against Scyan (Blampey et al.), another deep generative model used to classify cell populations in cytometry data, based on normalizing flows.

**Population discovery**: To test the model's ability to discover novel subpopulations, we mask all labels for TCD4 PeaReactive cells (TCD4pr) (non-canonical cells of interest) in the POISED dataset during training, aiming to see if the model is able to rediscover them without any supervision signal. In the second experience, we use the MRD dataset to identify blastic cells that were not used during the training of the model.

## RESULTS

**Classification performance**: MARVIN outperforms Scyan across all datasets and metrics. For the AML, BMMC, and POISED datasets, MARVIN achieves superior accuracy (0.99 vs 0.98, 0.99 vs 0.96, 0.95 vs 0.81), balanced accuracy (0.98 vs 0.89, 0.93 vs 0.80, 0.90 vs 0.78), and F1 scores (0.97 vs 0.83, 0.92 vs 0.75, 0.80 vs 0.58), respectively. The performance gap is most pronounced in the complex POISED dataset.

**Novel population discovery**:MARVIN successfully identifies the previously unlabeled TCD4pr population. Despite training without any labels for this population, the model correctly assigns these cells to a distinct cluster. Moreover, the model captures the biological dynamics of this population (emergence pattern following peanut stimulation, while still identifying memory cells at the unstimulated state). MARVIN correctly identifies unknown blast cell from the MRD dataset as it is unable to reconstruct this unknow population as a known population.

## SUMMARY/CONCLUSION

MARVIN provides a semi-supervised approach for cytometry data that effectively performs both classification and discovery tasks. The model demonstrates strong performance across multiple datasets and metrics. MARVIN successfully identifies unlabeled cell populations and characterizes their response patterns under different experimental conditions. This dual functionality makes it valuable for both research applications requiring discovery of novel populations and for routine clinical analysis such as MRD detection. We plan to apply this methodology to study immune activation patterns following vaccination.

**Keywords:** Myeloid Malignancies, Artificial Intelligence (AI), Diseases, Generative AI, Technology and Procedures, Deep Learning, Measurable Residual Disease , Bioinformatics, Lymphoid Malignancies

**Disclosure :** Adrien De Voeght: Johnson & Johnson, Consultancy (Includes expert testimony): Yes, Patents & Royalties: No, Ended employment in the past 24 months: No, Research Funding: No, Divested equity in a private or publicly-traded company in the past 24 months: No, Current equity holder in publicly-traded company: No, Current Employment: No, Current holder of stock options in a privately-held company: No, Current equity holder in private company: No, Honoraria: Yes, astellas pharma, Consultancy (Includes expert testimony): Yes, Patents & Royalties: No, Ended employment in the past 24 months: No, Research Funding: No, Divested equity in a private or publicly-traded company in the past 24 months: No, Current equity holder in publicly-traded company: No, Current Employment: No, Current holder of stock options in a privately-held company: No, Current equity holder in private company: No, Honoraria: Yes, Abbvie, Consultancy (Includes expert testimony): Yes, Patents & Royalties: No, Ended employment in the past 24 months: No, Research Funding: No, Divested equity in a private or publicly-traded company in the past 24 months: No, Current equity holder in publicly-traded company: No, Current Employment: No, Current holder of stock options in a privately-held company: No, Current equity holder in private company: No, Honoraria: Yes, Amgen, Consultancy (Includes expert testimony): Yes, Patents & Royalties: No, Ended employment in the past 24 months: No, Research Funding: No, Divested equity in a private or publicly-traded company in the past 24 months: No, Current equity holder in publicly-traded company: No, Current Employment: No, Current holder of stock options in a privately-held company: No, Current equity holder in private company: No, Honoraria: Yes, Servier, Consultancy (Includes expert testimony): Yes, Patents & Royalties: No, Ended employment in the past 24 months: No, Research Funding: No, Divested equity in a private or publicly-traded company in the past 24 months: No, Current equity holder in publicly-traded company: No, Current Employment: No, Current holder of stock options in a privately-held company: No, Current equity holder in private company: No, Honoraria: Yes, Fanny Bodart: None declared, Frederic Baron: None declared, Gilles Louppe: None declared