

Machine learning for integrating clinical and biological data

Author: Samuel Granjeaud ¹

¹ Centre de Recherche en Cancérologie de Marseille (CRCM), INSERM UMR1068, CNRS UMR7258, Aix-Marseille University UM105 and Paoli-Calmettes Institute, Marseille, France.

Keywords: Cytometry, Machine Learning, Data Analysis, Limits.

Abstract

Technological advances in flow cytometers and fluorochromes (or rare earth elements) make it possible to define panels targeting 30 to 50 markers, thereby identifying several dozen cell populations. In practice, multiparametric flow cytometry enables detailed profiling of dozens of biological samples by identifying the cell populations that make them up and then measuring their features, cell counts and the intensity of functional markers. By using multiple panels, dozens or even hundreds of features are measured for each biological sample, resulting in a large amount of information to analyze and synthesize.

The features measured by flow cytometry fall into two categories: counts, which are most often expressed as percentages, and intensities. These features have been analyzed for several decades, most often manually and in a ritualistic manner, which is labor-intensive and sometimes inaccurate. The shift toward analyzing a large number of features requires automated and more rigorous approaches.

The analysis of cell counts involves converting them into percentages, which is natural because the cell count of a cell population depends on the total count within a biological sample. However, there are constraints on percentages that are often overlooked. First, this information is bounded between 0% and 100%. Its distribution can be considered Gaussian for mean values between 20% and 80%, but it flattens out at the extremes, although transformations are possible to avoid this flattening (Tukey). In fact, common parametric statistical tests (Student's t-test, ANOVA) should be used with caution or replaced by nonparametric tests. Furthermore, when the percentage is statistically tested for each population one by one, a correction for multiple comparisons must be incorporated into the analysis to control for false positives. Finally, a percentage is calculated relative to a reference - typically the total number of cells - but the scientific question may call for other choices. These issues have been extensively studied in composition data analysis (CoDA) and single-cell RNA-seq. CoDA approaches break the mathematical dependence between populations within a biological sample and stabilize the variance of the transformed percentages by using a logarithmic transformation and by first replacing (imputing) zeros to avoid infinite values. scRNA-seq analyses (DESeq2, edgeR) rely directly on counts and their negative-binomial distribution, rather than on percentages, and zeros therefore do not require any specific handling. Percentages and their transformation remain important for visualizations, such as box plots or heatmaps (see the detailed protocol in Lohmann et al.).

Analyzing the intensity of functional markers (which are inherently a continuum: exhaustion, differentiation, etc.) typically requires the same transformation as that applied for their representation (bi-exp, asinh, etc.). Classical parametric tests can be applied provided that their assumptions are verified and the necessary correction for multiple testing is applied.

The previous analyses aim to identify features related to the scientific question and are univariate analyses (population by population). Machine learning approaches that perform classification or prediction are multivariate approaches that consider the set of features as an overall profile and aim to extract a signature related to the scientific question. This synthesis aims to avoid redundancy in the features and to assess the stability of the results. Machine

learning approaches have been developed progressively based on the ideas of their creators and are grounded in different lines of reasoning and choices, resulting in a broad spectrum of approaches to explore (LASSO, random forests, Support Vector Machines, etc.).

Their limitations include the small number of biological samples (per group or in total), which leads to a risk of instability and overfitting (strong reliance on training data, loss of generalization). The input data must usually be preprocessed (as mentioned earlier) and standardized so that the variation in the features can be analyzed on a common scale, regardless of whether the percentage is small or large, for example. To reduce the number of features comprising the signature, several strategies exist, such as the regularization applied in LASSO regression, or the competitive training of the STABL method (Hédou et al.), where the dataset competes with data representing pure noise. As a result, LASSO regression assigns a coefficient to each selected feature, but most other approaches are black boxes that lack interpretability. To address this shortcoming, methods such as SHAP (Shapley Additive Explanation) assess the importance of each feature by simulating different scenarios - for example, by observing how the prediction changes when a feature is present or absent.

Machine learning is now widely used in flow cytometry, as evidenced by reviews (Mocking et al., Wlosik et al.) and recent articles (Couckuyt et al., Emmaneel et al., Cross et al., Ashhurst et al.). Articles such as those by Greener et al. and Liu et al. provide a concise overview of these approaches from the perspective of an experimenter.

These approaches were used in the laboratory to identify immune biomarkers predictive of the efficacy of combination immunotherapy in patients with advanced cervical cancer (Bertucci et al.). Although combination immunotherapy improves outcomes, its overall efficacy is only 50%. It is therefore important to identify biomarkers that can guide treatment decisions. The study included 29 patients whose samples were analyzed by mass cytometry using three immune panels of 40 markers. Candidate biomarkers identified using conventional approaches (optSNE/UMAP + Phenograph, Citrus) were subjected to 11 distinct machine learning algorithms and an explainability method (SHAP). A cell population emerged from this set of results and was confirmed by conventional gating. This combination of predictive performance (machine learning) and biological transparency (explainability) constitutes a powerful form of cross-validation.

For univariate analyses, we developed Analycyte (Lohmann et al.), a Graphical User Interface (GUI) for exploratory and supervised feature analysis. This Web app allows to perform the typical analyses, i) Quality Control (QC): Assessing the relationship between features and experimental populations as well as detecting anomalies (e.g. batch effects), ii) Differential Abundance (DA): Identifying populations with changing abundance, iii) Differential State (DS): Identifying populations-marker combinations with changing marker intensity. Analycyte is compatible with concatenated Flow Cytometry Standard (FCS) file or Excel files (abundance and median fluorescence intensity (MFI)) with metadata.

For multivariate analyses (score, signature), we plan to integrate the Machine Learning approaches we already tested (Lasso, Random Forest, PLS...) for classification and survival, as Analycyte's architecture makes it easy to implement new analyses.

Analycyte enables experimenters and platforms to quickly obtain relevant post-clustering analyses that can be easily shared with collaborators (HTML documents). It is available free of charge as R packages or a Docker image that can run on the IFB cloud, and it supports FAIR (Findable, Accessible, Interoperable, Reusable) analyses.

References

Ashhurst TM, Marsh-Wakefield F, Putri GH, et al (2022) Integration, exploration, and analysis of high-dimensional single-cell cytometry data using Spectre. *Cytometry Part A* 101:237–253. <https://doi.org/10.1002/cyto.a.24350>

Bertucci A, Maby P, Chrétien AS, et al (in revision) Circulating exhausted TEMRA CD8+ T cells are associated with efficacy of combined aPD-L1, aCTLA-4 and metronomic chemotherapy in advanced cervical cancer

Couckuyt A, Van Gassen S, Emmaneel A, et al (2025) Unraveling genotype–phenotype associations and predictive modeling of outcome in acute myeloid leukemia. *Cytometry Part B: Clinical Cytometry* 108:366–377. <https://doi.org/10.1002/cyto.b.22230>

Cross CE, Brockman AA, Ihrie RA, Irish JM Velociraptor Machine Learning Quantifies Similarity to Known Cell Types and Matches Cells Across Flow and Imaging Cytometry Platforms. *Cytometry Part A*: <https://doi.org/10.1002/cyto.a.70066>

Emmaneel A, Neirinck J, Torres-Valle A, et al (2026) PIDgeon: An Explainable AI Model for Improved Flow Cytometry-Based Screening of Lymphoid Primary Immunodeficiencies. *Clin Chem* 72:601–612. <https://doi.org/10.1093/clinchem/hvaf190>

Greener JG, Kandathil SM, Moffat L, Jones DT (2022) A guide to machine learning for biologists. *Nat Rev Mol Cell Biol* 23:40–55. <https://doi.org/10.1038/s41580-021-00407-0>

Hédou J, Marić I, Bellan G, et al (2024) Discovery of sparse, reliable omic biomarkers with Stabl. *Nat Biotechnol* 42:1581–1593. <https://doi.org/10.1038/s41587-023-02033-x>

Lohmann E, Gorvel L, Granjeaud S (2025) Build heatmaps of cell frequencies and marker intensities for correct interpretations. 2025.02.10.637380 to be published in *Immunophenotyping, 2nd edition: Methods and Protocols* <https://www.biorxiv.org/content/10.1101/2025.02.10.637380v1>

Lohmann E, Mazuel A, Granjeaud S *Analycyte*: example and documentation <https://i-cyto.github.io/posts/240507-analycyte/>

Liu Y, Chen P-HC, Krause J, Peng L (2019) How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA* 322:1806–1816. <https://doi.org/10.1001/jama.2019.16489>

Mocking TR, van de Loosdrecht AA, Cloos J, Bachas C (2025) Applications of machine learning for immunophenotypic measurable residual disease assessment in acute myeloid leukemia. *HemaSphere* 9:e70138. <https://doi.org/10.1002/hem3.70138>

Tukey J (1977) *Exploratory Data Analysis*. Pearson, Reading, Mass

Wlosik J, Granjeaud S, Gorvel L, et al (2024) A beginner's guide to supervised analysis for mass cytometry data in cancer biology. *Cytometry Part A* 105:853–869. <https://doi.org/10.1002/cyto.a.24901>

Résumé

Les avancées technologiques sur les cytomètres et les fluochromes (ou métaux rares) permettent de définir des panels ciblant 30 à 50 marqueurs, identifiant ainsi plusieurs dizaines de populations. De fait, la cytométrie multi-paramétrique rend possible un profilage fin de plusieurs dizaines d'échantillons biologiques en déterminant les populations cellulaires qui les composent, puis en mesurant leurs caractéristiques, nombre de cellules et intensité de marqueurs fonctionnels. En utilisant plusieurs panels, plusieurs dizaines voire des centaines de caractéristiques sont mesurées sur chaque échantillon biologique, ce qui représente une quantité importante d'information à analyser et synthétiser.

Les caractéristiques mesurées en cytométrie sont de deux natures, d'une part des comptages, le plus souvent ramenés à des pourcentages, et d'autre part des intensités. L'analyse de ces caractéristiques est effectuée depuis plusieurs décennies, de façon le plus souvent manuelle et rituelle, ce qui est laborieux et parfois incorrect. Le passage à un grand nombre de caractéristiques nécessite des approches automatisées et plus rigoureuses.

L'analyse des comptages passe par une étape de transformation en pourcentage, ce qui est naturel parce que le comptage de cellules d'une population cellulaire dépend du comptage total au sein d'un échantillon biologique. Cependant, des contraintes souvent oubliées s'appliquent à un pourcentage. Tout d'abord, c'est une information bornée entre 0% et 100%. Sa distribution peut être considérée gaussienne pour des valeurs moyennes situées entre 20% et 80%, mais dans les zones extrêmes elle s'aplatit, bien que des transformations sont possibles pour éviter cet aplatissement (Tukey). De fait, les tests statistiques paramétriques courants (test de Student, ANOVA) doivent être employés avec précaution, ou remplacés par des tests non-paramétriques. De plus, lorsque le pourcentage est testé statistiquement pour chaque population, une à une, une correction de tests multiples doit impérativement être intégrée à l'analyse pour contrôler le nombre de faux positifs. Enfin, un pourcentage est calculé par rapport à une référence, le total de cellules en général, mais la question scientifique peut s'orienter vers d'autres choix. Ces problématiques ont été largement étudiées dans les analyses de données de composition (CoDA) et en single-cell RNA-seq. Les approches CoDA brisent la dépendance mathématique entre les populations au sein d'un échantillon biologique et stabilisent la variance des pourcentages transformés en utilisant une transformation logarithme et le remplacement (imputation) préalable des zéros pour éviter des valeurs infinies. Les analyses scRNA-seq (DESeq2, edgeR) s'appuient directement sur les comptages et leur distribution négative-binomiale, et non pas sur les pourcentages, les zéros ne nécessitant alors pas de traitement spécifique. Les pourcentages et leur transformation restent importants pour les représentations, telles que les box-plots ou les heatmaps (voir protocole détaillé dans Lohmann et al.).

L'analyse de l'intensité des marqueurs fonctionnels (par nature un continuum : exhaustion, différenciation...) nécessite typiquement la même transformation que celle qui est appliquée pour leur représentation (bi-exp, asinh...). Des tests paramétriques classiques peuvent être appliqués moyennant la vérification de leurs hypothèses et une indispensable correction de tests multiples.

Les analyses précédentes visent à identifier des caractéristiques liées à la question scientifique et sont des analyses univariées (population par population). Les approches de Machine Learning réalisant une classification ou une prédiction sont des approches multivariées qui considèrent l'ensemble des caractéristiques comme un profil global et visent à en extraire une signature liée à la question scientifique. Cette synthèse veut éviter la redondance des caractéristiques et évaluer la stabilité du résultat. Les approches de Machine Learning ont été bâties de façon progressive sur les idées de leur concepteur et

reposent sur des réflexions et choix différents, ce qui donne un large spectre d'approches à exploiter (LASSO, forêts aléatoires, Support Vecteur Machine...).

Leurs limites sont la faiblesse du nombre d'échantillons biologiques (par groupe ou au total), qui entraîne un risque d'instabilité et de sur-apprentissage (forte dépendance aux données d'entraînement, perte de généralisation). Les données d'entrée doivent être le plus souvent pré-transformées (comme précédemment) et standardisées afin que la variation des caractéristiques soit analysée sur une échelle commune, que le pourcentage soit petit ou grand par exemple. Pour réduire le nombre de caractéristiques constituant la signature, plusieurs stratégies existent telles que la pénalisation appliquée dans la régression LASSO, ou l'entraînement compétitif de la méthode STABL (Hédou et al.), où le jeu de données est mis en compétition avec des données représentant purement du bruit. En résultat, la régression LASSO assigne un coefficient à chaque caractéristique retenue, mais la plupart des autres approches sont des boîtes noires qui manquent d'explicabilité. Pour combler ce manque, des méthodes comme SHAP (Shapley additive explanation) évaluent l'importance de chaque caractéristique en simulant différents scénarios, par exemple en observant ce qui change dans la prédiction quand une caractéristique est présente ou absente.

L'utilisation du Machine Learning est fréquente maintenant en cytométrie comme en témoignent les revues (Mocking et al., Wlosik et al.) et les articles récents (Couckuyt et al., Emmaneel et al., Cross et al., Ashhurst et al.). Des articles (Greener et al., Liu et al.) permettent d'appréhender les approches de façon synthétique avec le regard d'un expérimentateur.

Ces approches ont été utilisées au laboratoire pour identifier des biomarqueurs immunitaires prédictifs de l'efficacité de l'immunothérapie combinée chez les patientes atteintes de cancer du col utérin avancé (Bertucci et al.). Bien que l'immunothérapie combinée améliore les résultats, son efficacité globale n'est que de 50%. Il est donc important d'identifier des biomarqueurs susceptibles de guider la prescription thérapeutique. L'étude a porté sur 29 patientes dont les échantillons ont été analysés en cytométrie de masse avec 3 panels immuns de 40 marqueurs. Les biomarqueurs candidats identifiés par des approches classiques (optSNE/UMAP + Phenograph, Citrus) ont été soumis à 11 algorithmes distincts de Machine Learning et une méthode d'explicabilité (SHAP). Une population cellulaire est ressortie de cet ensemble de résultats et a été confirmée gating conventionnel. Cette combinaison de performance prédictive (machine learning) et de transparence biologique (explicabilité) constitue une validation croisée puissante.

Pour les analyses univariées, nous avons développé Analycyte (Lohmann et al.), une interface utilisateur graphique (GUI) destinée à l'analyse exploratoire et supervisée des caractéristiques. Cette application Web permet d'effectuer les analyses classiques suivantes : i) Contrôle qualité (CQ) : évaluation de la relation entre les caractéristiques et les populations expérimentales, ainsi que détection des anomalies (par exemple, les effets de lot) ; ii) Abondance différentielle (DA) : l'identification des populations dont l'abondance varie, iii) l'état différentiel (DS) : l'identification des combinaisons population-marqueur dont l'intensité du marqueur varie. Analycyte est compatible avec les fichiers FCS (Flow Cytometry Standard) concaténés ou les fichiers Excel (abondance et intensité de fluorescence médiane (MFI)) accompagnés de métadonnées.

Pour les analyses multivariées (score, signature), nous prévoyons d'intégrer les approches d'apprentissage automatique que nous avons déjà testées (Lasso, Random Forest, PLS...) pour la classification et la survie, car l'architecture d'Analycyte facilite la mise en œuvre de nouvelles analyses.

Analycyte permet aux chercheurs et aux plateformes d'obtenir rapidement des analyses pertinentes post-regroupement, qui peuvent être facilement partagées avec des collaborateurs (documents HTML). Il est disponible gratuitement sous forme de paquets R ou d'image Docker pouvant s'exécuter sur le cloud de l'IFB, et prend en charge les analyses FAIR (Findable, Accessible, Interoperable, Reusable).

Références

Ashhurst TM, Marsh-Wakefield F, Putri GH, et al (2022) Integration, exploration, and analysis of high-dimensional single-cell cytometry data using Spectre. *Cytometry Part A* 101:237–253. <https://doi.org/10.1002/cyto.a.24350>

Bertucci A, Maby P, Chrétien AS, et al (in revision) Circulating exhausted TEMRA CD8+ T cells are associated with efficacy of combined aPD-L1, aCTLA-4 and metronomic chemotherapy in advanced cervical cancer

Couckuyt A, Van Gassen S, Emmaneel A, et al (2025) Unraveling genotype–phenotype associations and predictive modeling of outcome in acute myeloid leukemia. *Cytometry Part B: Clinical Cytometry* 108:366–377. <https://doi.org/10.1002/cyto.b.22230>

Cross CE, Brockman AA, Ihrie RA, Irish JM Velociraptor Machine Learning Quantifies Similarity to Known Cell Types and Matches Cells Across Flow and Imaging Cytometry Platforms. *Cytometry Part A*: <https://doi.org/10.1002/cyto.a.70066>

Emmaneel A, Neirinck J, Torres-Valle A, et al (2026) PIDgeon: An Explainable AI Model for Improved Flow Cytometry-Based Screening of Lymphoid Primary Immunodeficiencies. *Clin Chem* 72:601–612. <https://doi.org/10.1093/clinchem/hvaf190>

Greener JG, Kandathil SM, Moffat L, Jones DT (2022) A guide to machine learning for biologists. *Nat Rev Mol Cell Biol* 23:40–55. <https://doi.org/10.1038/s41580-021-00407-0>

Hédou J, Marić I, Bellan G, et al (2024) Discovery of sparse, reliable omic biomarkers with Stabl. *Nat Biotechnol* 42:1581–1593. <https://doi.org/10.1038/s41587-023-02033-x>

Lohmann E, Gorvel L, Granjeaud S (2025) Build heatmaps of cell frequencies and marker intensities for correct interpretations. 2025.02.10.637380 à paraître dans *Immunophenotyping*, 2nd edition: Methods and Protocols <https://www.biorxiv.org/content/10.1101/2025.02.10.637380v1>

Lohmann E, Mazuel A, Granjeaud S Analycyte: example and documentation <https://i-cyto.github.io/posts/240507-analycyte/>

Liu Y, Chen P-HC, Krause J, Peng L (2019) How to Read Articles That Use Machine Learning: Users' Guides to the Medical Literature. *JAMA* 322:1806–1816. <https://doi.org/10.1001/jama.2019.16489>

Mocking TR, van de Loosdrecht AA, Cloos J, Bachas C (2025) Applications of machine learning for immunophenotypic measurable residual disease assessment in acute myeloid leukemia. *HemaSphere* 9:e70138. <https://doi.org/10.1002/hem3.70138>

Tukey J (1977) *Exploratory Data Analysis*. Pearson, Reading, Mass

Wlosik J, Granjeaud S, Gorvel L, et al (2024) A beginner's guide to supervised analysis for mass cytometry data in cancer biology. *Cytometry Part A* 105:853–869. <https://doi.org/10.1002/cyto.a.24901>