Received: 20 August 2021 | Revised: 2 February 2022 | Accepted: 28 February 2022

DOI: 10.1002/cyto.b.22065

ORIGINAL ARTICLE

CLINICAL CYTOMETRY WILEY

Prediction of standard cell types and functional markers from textual descriptions of flow cytometry gating definitions using machine learning

Raul Rodriguez-Esteban[1] | José Duarte[1] | Priscila C. Teixeira[1] | Fabien Richard[1] | Svetlana Koltsova[2] | W. Venus So[3]

[1]Roche Pharmaceutical Research and Early Development, Roche Innovation Center Basel, Basel, Switzerland

[2]Curation Department, Rancho BioSciences LLC, San Diego, California, USA

[3]Roche Pharmaceutical Research and Early Development, Roche Innovation Center New York, New York, USA

Correspondence
Raul Rodriguez-Esteban, Roche Pharmaceutical Research and Early Development, Roche Innovation Center Basel, Grenzacherstrasse 124, 4070, Basel, Switzerland.
Email: raul.rodriguez-esteban@roche.com

Abstract

Background: A key step in clinical flow cytometry data analysis is gating, which involves the identification of cell populations. The process of gating produces a set of reportable results, which are typically described by gating definitions. The non-standardized, non-interpreted nature of gating definitions represents a hurdle for data interpretation and data sharing across and within organizations. Interpreting and standardizing gating definitions for subsequent analysis of gating results requires a curation effort from experts. Machine learning approaches have the potential to help in this process by predicting expert annotations associated with gating definitions.

Methods: We created a gold-standard dataset by manually annotating thousands of gating definitions with cell type and functional marker annotations. We used this dataset to train and test a machine learning pipeline able to predict standard cell types and functional marker genes associated with gating definitions.

Results: The machine learning pipeline predicted annotations with high accuracy for both cell types and functional marker genes. Accuracy was lower for gating definitions from assays belonging to laboratories from which limited or no prior data was available in the training. Manual error review ensured that resulting predicted annotations could be reused subsequently as additional gold-standard training data.

Conclusions: Machine learning methods are able to consistently predict annotations associated with gating definitions from flow cytometry assays. However, a hybrid automatic and manual annotation workflow would be recommended to achieve optimal results.

KEYWORDS
automatic text annotation, flow cytometry gating, gating definitions, assay annotation

1 | INTRODUCTION

Flow cytometry enables high content analysis of cell populations from heterogeneous samples through the identification of surface and intracellular antigen expression using fluorescent-labeled molecular probes (Chattopadhyay et al., 2008). It can provide insights in applications such as the identification of disease biomarkers, immune regulatory mechanisms, and cellular signaling. As such, flow cytometry is an important tool in drug discovery and development in areas such as biomarker discovery, receptor occupancy and target engagement assays, and target-based and phenotypic screenings (Edwards & Sklar, 2015; Gedye et al., 2014; Moulard & Ozoux, 2016).

Recent years have seen tremendous development in multiplexing capabilities of flow cytometry instrumentation, namely with the development of full spectrum flow cytometry (Nolan et al., 2013; Robinson, 2019). This innovation has reached the clinical space and is implemented in high parameter flow cytometry assays in multi-center clinical trials. Such trials generate data from hundreds to thousands of samples across multiple flow cytometry assays that are capable of reporting on hundreds to thousands of different reportable results ("reportables"). Leveraging these new capabilities, for instance, pharmaceutical companies employ an evolving mix of flow cytometry assays during the drug project life cycle, stemming from a range of internally-developed assays and potentially multiple external laboratories.

The sharing of the biomarker data produced by these assays enables its reuse, reanalysis, and reproducibility (Bhattacharya et al., 2018). However, effective sharing following best practices, such as those exemplified by the FAIR data sharing principles (Wilkinson et al., 2016), presents numerous challenges, such as those deriving from differences in sample management, instrumental setup, and data analysis (Finak et al., 2016; Maecker et al., 2010; Montante & Brinkman, 2019). Although much attention has been given to the harmonization and alignment of flow cytometry instruments in multi-center trials (Finak et al., 2016; Jamin et al., 2016; Larbi, 2017; White et al., 2015), there is still no guidance and/or tools for the standardization of flow cytometry data analysis and harmonization, such as the management of an assay's multiple reportables. Moreover, data inconsistencies and errors can make cross-study analysis particularly difficult to execute. These factors, among others, increase the burden of deploying high parameter flow cytometry in the clinic.

The particular case of validated flow cytometry assays from clinical research organizations (CRO) is worth highlighting for its complexity. Here, scientists are faced with complex assay development setups that include the creation of reportables for each new validated assay. Due to the increasing complexity of biomarker strategies in early phase clinical trials in the last few decades (Califf, 2018; Freidlin et al., 2010), it is becoming ever more common that target-specific assays have to be generated for each new molecule being assessed in the clinic (e.g., receptor occupancy assays), which, in turn, means that scientists face an increasing demand for annotating the new reportables that are being generated by the CROs.

These reportables often have strong limitations in their format and length due to the vendor's database setup or existing data model that supports data transfer from vendor to sponsor. Moreover, different CROs have different rules on how to create reportables, and some have no rules at all. Finally, these reportables can show a mix of multiple properties of what is being measured (cell type of interest, functional marker of interest, reference cell population, unit of measurement, etc.). In the past, scientists could easily decipher these reportables because flow cytometry in clinical trials would usually report only major lineage populations and assay complexity was small (e.g., CD45+CD3+CD8+(%CD3+)). With the introduction of high parameter instrumentation in the clinic, assays are now able to report on deep immunophenotyping and scientists are no longer faced with a list of a few dozen relatively simple reportables but with lists of potentially hundred to thousands of complex ones (e.g., CD45+CD3+CD8+CD45RA+CD197+CD95+CD28+ (%CD8+N)).

These complex flow cytometry datasets that can comprise dozens of different assays generated by multiple CROs throughout multiple clinical trials are a rich source of information for reverse translation efforts that are being carried out throughout the pharmaceutical industry. Standardization and automation mechanisms for flow cytometry reportables annotation metadata are critical factors for the success and simplification of said efforts.

The focus of this study is on the way reportables are described, which are typically represented using unstructured text strings, sometimes referred to as "gating definitions," that comprise relevant markers and other information about the assay. Due to a lack of widespread standards, gating definitions can be written in multiple ways, which is recognized as an obstacle for data sharing (Overton et al., 2019). For example, multiple gating definitions can be derived from a single gating hierarchy and, therefore, knowledge of the gating hierarchy is insufficient to characterize a gating definition. Integration of gating definitions from flow cytometry datasets is typically done through manual curation, with its associated perils, such as curation inconsistencies, drift and errors (Rodriguez-Esteban, 2015).

Overton et al. (2019) introduced an approach to check the validity and consistency of gating definitions for 4388 gating definitions produced by a set of 28 academic centers. Their approach leveraged ontology mapping and, in particular, the Cell Ontology (CL) (Diehl et al., 2016) and the Protein Ontology (PRO) (Chen et al., 2020; Natale et al., 2011). It involved, among other steps, mapping gating definitions to marker gene names and intensity levels using a rule-based method. As stated by the authors, however, pure rule-based approaches have shortcomings in dealing with textual ambiguity. Owing to incomplete ontologies, ontology mapping can lead to false negatives due to unmatched relevant concepts. Additionally, rule-based methods can struggle to capture complex relations between elements of the text.

In this study, we explored a related problem to that reported in Overton et al. (2019). We studied the feasibility of predicting cell types and functional markers associated with gating definitions with the help of machine learning (ML). Functional markers are markers that provide additional properties (e.g., proliferation and exhaustion status) but are not needed to define the cell types of interest in a particular assay. The cell types associated with each gating definition in an assay and the presence or absence of specific functional markers are of key interest for analyses concerning flow cytometry data because the annotation of gating definitions with standard concepts enables data integration and re-use. To tackle this problem, we applied a supervised ML approach. ML approaches for handling unstructured text have long been deployed in pharmaceutical research and development, which involves the mining of large amounts of textual information (Rodriguez-Esteban, 2016). In particular, mapping and classification of unstructured text is an area in which ML algorithms for text mining have already shown their utility (Rodriguez-Esteban, 2019). Examples of state-of-the-art ML algorithms used for text classification are based on deep learning neural networks (González-Carvajal & Garrido-Merchán, 2005; Minaee et al., 2021). However, traditional algorithms have been shown to outperform deep learning algorithms in tasks with low dimensionality, such as tabular classification (Kadra et al., 2021; Shwartz-Ziv & Armon, 2022). The selection and fine-tuning of such algorithms has often been done manually but automatic selection pipelines (autoML) (He et al., 2021) have been proposed as an unbiased alternative that is able to systematically explore different hyperparameters and configurations. Thus, in this paper, we wanted to explore the feasibility of automatically identifying cell types and functional markers from gating definitions using an ML algorithm selected through an automatic pipeline.

2 | METHODS

The dataset for this study comprised 4849 gating definitions from 36 assay panels belonging to four different laboratories. Despite differences between the assays, some gating definitions were identical. After deduplication, we had a total of 3045 unique gating definitions available.

Most of the unique gating definitions ($n = 3043$) were interpreted by scientific experts, which annotated them with 117 unique cell types and 70 unique functional markers. An example of annotation of interpreted cell type and functional marker can be seen in Table 1. These annotations initially lacked some consistency. That is, the same cell type or functional marker was written in different ways by different experts. To increase consistency of annotation, annotated cell types were mapped to an internal Roche cell type terminology which integrates domain experts' feedback and multiple public ontologies including CL, BRENDA Tissue Ontology, SNOMED, NCI Thesaurus and MeSH; and is hosted by the Roche Terminology System (RTS), which is an internally-developed platform for the management and distribution of highly curated terminologies currently covering around 130,000 concept entries, and which is completely built on a semantic technology stack providing uniform resource identifiers (URIs) to support data FAIRification at scale across all Roche functions and sites. Very briefly, the creation of a cell type concept in RTS depends on experimental evidence showing that the said cell type has cellular functions different from other existing cell types. The mapping to the RTS terminologies involved manual expert curation and, additionally, rule-based automated quality control. For example, annotated marker gene names were harmonized to CD names where available. After the harmonization, we had annotations corresponding to 56 unique cell types and 62 unique functional markers, which became the target variables for the ML algorithm.

TABLE 1 Examples of gating definitions mapped to cell types and markers

Gating definition	Cell type	Functional marker
CD3+CD4+CD25_APC MFI	Lymphocyte T, CD4-positive	CD25
CD8+PD-1 Tcell MESFNaHepLDTCL	Lymphocyte T, CD8-positive	CD279
Median CTLA4 in CD19	Lymphocyte B	CD152
%4-1BB in CD16	Natural killer cell	CD137

Generalizability and reproducibility were emphasized in building the overall ML prediction workflow. Gating definitions were preprocessed by transforming them to lowercase, eliminating non-ASCII characters and most non-alphanumeric characters. Following Overton et al. (2019), a simple set of rules was applied to split ("tokenize") gating definitions into units by identifying "separator" elements. These units ("tokens") often corresponded to individual gates (e.g., "CD3+CD4+CD25+" was split into the tokens CD3, CD4, and CD25). Marker intensity definitions (e.g., plus and minus signs next to individual gates, such as + in "CD3+"), where they existed, were extracted for each token.

The dataset was divided into training (80%) and test (20%) sets. Features for ML were based on all unique tokens produced by tokenization of the training dataset. These features were then matched to all gating definitions in the training and test sets to produce, respectively, the training and test feature values. Matches were not allowed when there were numerical boundaries around the match (e.g., the feature 45ra matched the gating definition "CD45RA+" but the feature cd4 did not). Marker intensity definitions were used to further refine feature values (e.g., a minus sign led to a feature value of -1).

Ontology matching was applied to gating definitions to reduce feature-set cardinality. For this purpose, the PRO release 62.0 in OWL format, which included 331,920 terms, was downloaded from the Protein Ontology Consortium site (proconsortium.org).

An ML pipeline was chosen with the aid of the TPOT autoML (automated ML) library. AutoML algorithms can help the end-to-end selection of an optimal pipeline of preprocessors, feature constructors, feature selectors, ML models and hyperparameter optimization for solving an ML task (He et al., 2021). The TPOT autoML library for classification (Olson & Moore, 2016; v. 0.11.7) selects a model from a list that, in its default configuration, includes Gaussian naïve Bayes, Bernoulli naïve Bayes, multinomial naïve Bayes, decision tree, extra trees, random forest, gradient boosting, K-nearest neighbors, linear support vector machine, logistic regression, extreme gradient boosting, stochastic gradient descent, and multi-layer perceptron. The ML pipeline was selected by running the TPOT autoML algorithm on the training set. The selected pipeline was primarily evaluated in the test set, which had not been seen by the autoML algorithm. Secondarily, it was evaluated by 10-fold cross-validation on the entire dataset.

All code and data created for this study are available at: https://github.com/raroes/prediction-flow-cytometry-gating-definitions

3 | RESULTS

A total of 3043 gating definitions were manually annotated by scientific experts and harmonized to 56 unique cell types and 62 unique

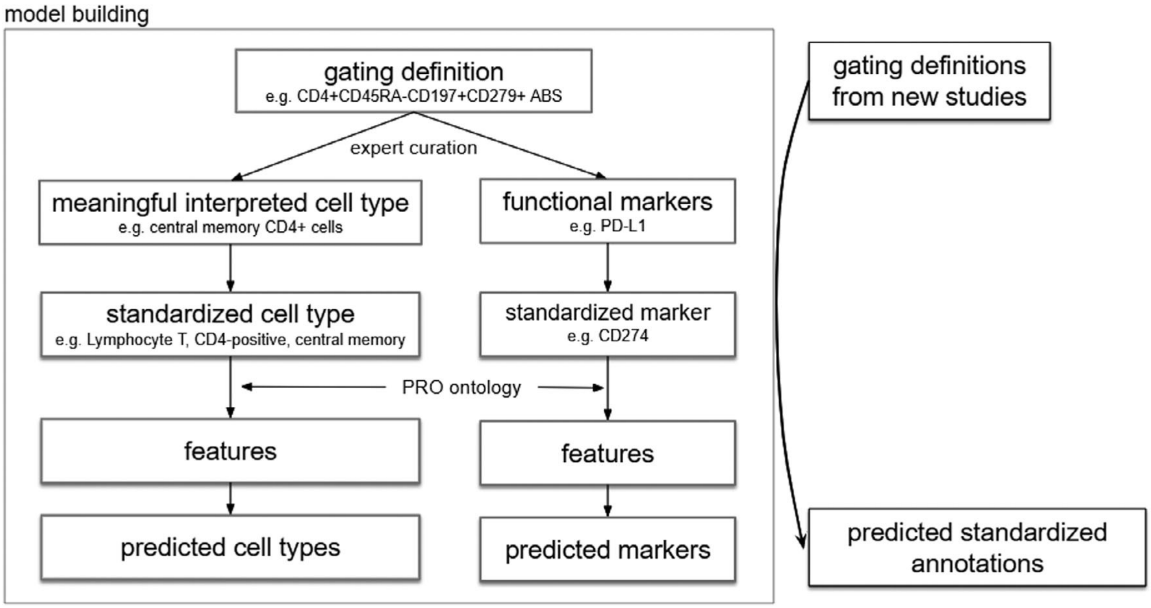
model building
gating definition
e.g. CD4+CD45RA-CD197+CD279+ ABS
expert curation
meaningful interpreted cell type
e.g. central memory CD4+ cells
functional markers
e.g. PD-L1
standardized cell type
e.g. Lymphocyte T, CD4-positive, central memory
standardized marker
e.g. CD274
PRO ontology
features
features
predicted cell types
predicted markers
gating definitions from new studies
predicted standardized annotations

FIGURE 1 Prediction workflow for gating definitions mapped to ontology cell types. Gating definitions are manually mapped to meaningful cell types and functional markers. Cell types and functional markers are chosen manually by experts based on their interpretation of the definition. Afterward, these are mapped to standardized cell types and markers, with the aid of the PRO ontology, to serve as prediction targets of the ML algorithm. ML, machine learning

functional markers (Table 1). Using this dataset, we developed a ML-based prediction workflow to predict the cell type and functional marker annotation associated with each gating definition, that is, to solve a 56-class and a 62-class classification problem, respectively (Figure 1 for the workflow schematic).

For cell type annotation prediction, the data was split into training and test datasets. Based on the gating definitions in the training dataset, a total of 281 features were created through data pre-processing steps (see Section 2). Feature values were extracted for both training and test datasets to feed the ML pipeline. An ML pipeline was selected and optimized by the TPOT autoML (automated machine learning) algorithm. This pipeline was based on a stacking architecture composed of a random forest classifier and a logistic regression classifier. Using this pipeline, prediction accuracy on the test set was 97.2%. The median area under the curve of the receiver operating characteristic (AUROC) for each class was 0.999 and the average AUROC was 0.95 ± 0.12. Low performance for classes with few available gating definitions in the training lowered the average AUROC. The histogram distribution of AUROC values for all cell type classes is shown in Figure 2. The overall 10-fold cross-validation accuracy for the ML pipeline was 94.2%.

Due to the way manually-curated cell type annotations were written, these could be split into segments separated by commas. For example, "Lymphocyte T, CD8-positive, regulatory" could be split into three segments: "Lymphocyte T," "CD8-positive" and "regulatory" (Table 1). An error analysis showed that, out of 177 classification errors made by the ML algorithm, 124 (70.1% of all errors) corresponded to discrepancies with the third segment of the annotations (sub-subcategories). For example, predicting "Lymphocyte T, CD4-positive" when the actual class was "Lymphocyte T, CD4-positive, *naive*." Most predictions were, on the other hand, correct with respect to the first segment (main category), which corresponded to the broader cell type (e.g., monocyte, neutrophil, T-cell), except in 34 cases (19.2% of all errors). Table 2 shows a summary of error types for cell type annotation predictions. A large share of errors was due to predictions that were not specific enough (39%).

Data availability by source (i.e., external laboratory or internal assay provider) varied widely, as can be seen in Table 3. To test the ability of an ML pipeline trained on data from one source to successfully make predictions on test data from another source (i.e., transfer learning), we performed several experiments. First, we tested a pipeline on data from a single source after being trained on the rest of the sources. Results of this experiment can be seen in the first column of Table 4 and indicate that lack of same-source data in the training set had a strong negative impact on ML pipeline performance. The second experiment involved mixing different amounts of same-source and other-source data in the training set. As can be seen in Table 4, the addition of more data to the training set, whether same-source or other-source, generally improved prediction accuracy. It can also be seen that including even as little as 10% of same-source data significantly increased prediction accuracy in most cases.

We performed an analogous analysis with the prediction of functional marker annotations. We manually labeled 3038 gating definitions with functional marker annotations corresponding to 62 unique marker names (Table 1). This dataset was then used for creating, training and testing an ML pipeline as previously described. An ML pipeline based on logistic regression with L2 regularization was selected by the TPOT autoML algorithm. The accuracy of this ML pipeline was

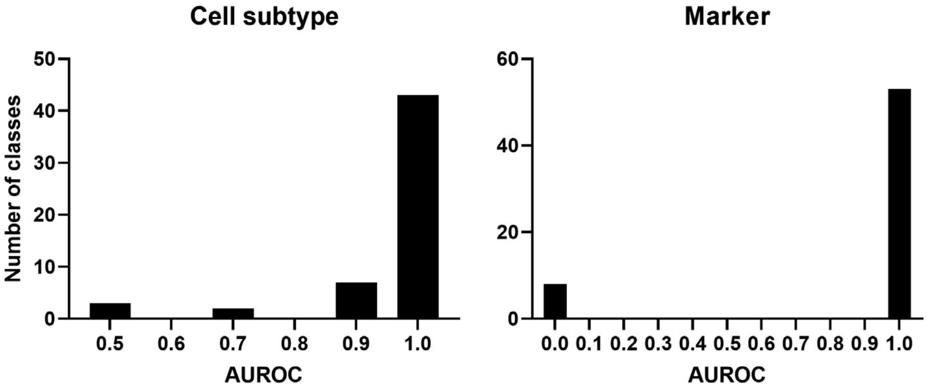

FIGURE 2 Histogram of AUROC values associated with the classification of each cell type and functional marker class. The abscissa is labeled with the upper bound of the histogram intervals. AUROC: area under the curve of the receiver operating characteristic

TABLE 2 Error analysis of cell type predictions

Error type	%	Example prediction	Example actual value
Prediction too specific	9	Lymphocyte T, CD4-positive, regulatory	Lymphocyte T, CD4-positive
Prediction too general	39	Myeloid dendritic cell	Myeloid dendritic cells, mDC1
Main category error	19	Immature neutrophil	Intermediate monocyte
Subcategory error	7	Lymphocyte T, helper type 1	Lymphocyte T, CD4-positive, regulatory
Sub-subcategory error	27	Lymphocyte T, CD4-positive, central memory	Lymphocyte T, CD4-positive, effector memory

Note: For each error type an example is shown of an erroneous prediction. The largest share of errors was due to predictions that were not specific enough (39%).

TABLE 3 Number and percentage of curated gating definitions available by source (i.e., external laboratory or internal assay provider)

Source	%	*n*
Internal	3.0	92
Laboratory 1	83.8	2551
Laboratory 2	6.0	184
Laboratory 3	7.1	216

98.5% on the test set. The median AUROC was 1.00 and the AUROC average was 0.87 ± 0.32 (Figure 2). The overall 10-fold cross-validation accuracy was 95.0%.

We then carried out the same experiments that we had performed with cell type annotation prediction in order to test the performance of the ML pipeline when trained on data from a mix of sources (Table 5). Similarly to the case with cell types, an increase in training data availability, whether same-source or other-source, led to greater accuracy (with one exception) and small amounts of same-source data improved performance considerably. Table 6 shows an error analysis for marker annotation predictions. The largest share of errors was due to predictions that did not specify a marker (52%).

Because gene names are highly ambiguous (Liu et al., 2006), we also explored following Overton et al. (2019) by using a gene name ontology (PRO) to identify features that could have been derived from different synonyms for the same gene name (see workflow in Figure 1). A total of 16 features identified in the training set that corresponded to synonyms from eight genes were merged (e.g., the features corresponding to the gene name synonyms ICOS and CD278 were merged into one feature). This, however, did not improve performance for either cell type or marker classification.

4 | DISCUSSION AND CONCLUSIONS

This study showed the feasibility of using ML algorithms for annotating flow cytometry gating definitions with standardized cell types and functional markers, thereby enabling the interpretation and integration of data provided by different assays deployed in multi-center studies. More accurate and efficient data integration increases the value of data and enhances its ability to generate clinical and biological insights. The ML algorithms chosen in this study can easily be re-trained as additional curated gating definitions are added to the training data.

Overall, the annotation of functional markers showed better performance than the annotation of cell types, which is possibly associated with differences in task complexity (Rodriguez-Esteban & Loging, 2013). While the training data that best helped predicting annotations belonged typically to assays developed by the same laboratory, we showed that the inclusion of gating definitions from assays belonging to other laboratories in the training data generally provided additional predictive power, thereby confirming the value of transfer learning.

Cell type annotation errors were most frequent in fine-grained cell subtypes, as the main cell type was usually correctly predicted. As would be expected, annotating cell types and functional markers for which few examples were available in the training data was a challenge for the ML pipeline. This could be seen in the decreased

TABLE 4 Accuracy of the prediction pipeline for cell types when tested on single-source data and trained on different sets of sources

Training data	% of existing same-source	0	10		50	
	% of existing other-sources	100	0	100	0	100
Test data	Internal	66.3	25.3	83.1	89.1	91.3
	Laboratory 1	26.9	84.4	88.5	98.1	98.1
	Laboratory 2	58.2	47.6	63.3	62.0	75.0
	Laboratory 3	24.5	42.1	59.0	87.0	92.6

Note: The first column of results corresponds to a pipeline tested on 100% of the data available from a single source (same-source) and trained on data from the rest of the sources (other-source). The second column of results corresponds to an algorithm trained on 10% of same-source data plus 0% or 100% of other-source data. The third column of results corresponds to an algorithm trained on 50% same-source data plus 0% or 100% of other-source data.

TABLE 5 Accuracy of the prediction pipeline for markers when tested on single-source data and trained on different sources

Training data	% of existing same-source	0	10		50	
	% of existing other-sources	100	0	100	0	100
Test data	Internal	26.1	38.6	74.7	91.3	97.8
	Laboratory 1	15.5	83.1	83.8	98.8	96.8
	Laboratory 2	81.5	70.5	86.7	94.6	98.9
	Laboratory 3	81.9	74.9	97.4	92.6	98.1

Note: The first column of results corresponds to an algorithm tested on 100% of the data available from a single source (same-source) and trained on data from the rest of the sources (other-source). The second column of results corresponds to an algorithm trained on 10% of same-source data plus 0% or 100% of other-source data. The third column of results corresponds to an algorithm trained on 50% same-source data plus 0% or 100% of other-source data.

TABLE 6 Error analysis of cell type predictions

Error type	%	Example prediction	Example actual value
No marker predicted	52	No marker	CD25
No marker available	30	HLA-DR	No marker
Partial marker error	4	CD366	CD279+CD366+
Full marker error	14	CD274	CD279

Note: For each error type an example is shown of an erroneous prediction. The largest share of errors was due to predictions that did not specify a marker (52%).

AUROC for classes with low number of samples. Additionally, we observed that the annotation of gating definitions from assays belonging to laboratories for which there was no training data could lead to poor performance due to differences in the way gating definitions are written across laboratories. This can be addressed, as shown in our experiments, by curating a small set of representative gating definitions, so that the algorithm can learn to recognize the feature patterns that define data from a never-before-seen laboratory. Thus, while we have noted that "the more data the better," manually-curated data can, nonetheless, be gathered strategically to increase its representativeness and improve the performance of the ML pipeline at low cost.

Additional challenges to our ML approach include those typical from operationalization of an ML algorithm (Mäkinen et al., 2021), such as tracking of ML model and dataset versions, as well as maintaining consistency in the quality of manual annotations. The ML algorithm itself can help in identifying consistency errors in manual annotations if an error analysis is performed on its predictions. Based on our practice, the output of the ML algorithm should be manually checked, which ensures high quality in the final output with minimal manual work. This output, in turn, can become additional high quality training data.

A clear advantage of a purely ML approach over a rule-based approach is that it does not depend on the currency, comprehensiveness or quality of the rules or ontologies used in the latter. However, the use of an ML approach does not preclude the inclusion of rules or ontologies. In fact, a mixed approach in which rules or ontologies were used to engineer features may improve performance. In our case, we tested the widely used ontology PRO to enhance our pipeline by normalizing features derived from synonyms corresponding to the same genes. This normalization did not, however, lead to an improvement in performance, perhaps due to the limited number of ambiguous gene synonyms identified.

ACKNOWLEDGMENTS

We would like to thank our managers Barbara Endler-Jobst and James Cai for their support. We would also like to thank our colleagues Iva Lelios, Fabian Junker, Nima Salimi, Yelena Budovskaya and Enrique Gomez for their help with assay annotation.

CONFLICT OF INTEREST

The authors declare no potential conflict of interest.

ORCID

Raul Rodriguez-Esteban https://orcid.org/0000-0002-9494-9609

REFERENCES

Bhattacharya, S., Dunn, P., Thomas, C. G., Smith, B., Schaefer, H., Chen, J., Hu, Z., Zalocusky, K. A., Shankar, R. D., Shen-Orr, S. S., Thomson, E., Wiser, J., & Butte, A. J. (2018). ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Scientific Data*, *27*(5), 180015.

Califf, R. M. (2018). Biomarker definitions and their applications. *Experimental Biology and Medicine*, *243*(3), 213–221.

Chattopadhyay, P. K., Hogerkorp, C. M., & Roederer, M. (2008). A chromatic explosion: The development and future of multiparameter flow cytometry. *Immunology*, *125*(4), 441–449.

Chen, C., Huang, H., Ross, K. E., Cowart, J. E., Arighi, C. N., Wu, C. H., & Natale, D. A. (2020). Protein ontology on the semantic web for knowledge discovery. *Scientific Data*, *7*(1), 337.

Diehl, A. D., Meehan, T. F., Bradford, Y. M., Brush, M. H., Dahdul, W. M., Dougall, D. S., He, Y., Osumi-Sutherland, D., Ruttenberg, A., Sarntivijai, S., Van Slyke, C. E., Vasilevsky, N. A., Haendel, M. A., Blake, J. A., & Mungall, C. J. (2016). The Cell Ontology 2016: Enhanced content, modularization, and ontology interoperability. *Journal of Biomedical Semantics*, *7*(1), 44.

Edwards, B. S., & Sklar, L. A. (2015). Flow cytometry: Impact on early drug discovery. *Journal of Biomolecular Screening*, *20*(6), 689–707.

Finak, G., Langweiler, M., Jaimes, M., Malek, M., Taghiyar, J., Korin, Y., Raddassi, K., Devine, L., Obermoser, G., Pekalski, M. L., Pontikos, N., Diaz, A., Heck, S., Villanova, F., Terrazzini, N., Kern, F., Qian, Y., Stanton, R., Wang, K., … McCoy, J. P. (2016). Standardizing flow cytometry immunophenotyping analysis from the human Immunophenotyping consortium. *Scientific Reports*, *10*(6), 20686.

Freidlin, B., McShane, L. M., & Korn, E. L. (2010). Randomized clinical trials with biomarkers: Design issues. *Journal of the National Cancer Institute*, *102*(3), 152–160.

Gedye, C. A., Hussain, A., Paterson, J., Smrke, A., Saini, H., Sirskyj, D., Pereira, K., Lobo, N., Stewart, J., Go, C., Ho, J., Medrano, M., Hyatt, E., Yuan, J., Lauriault, S., Meyer, M., Kondratyev, M., van den Beucken, T., Jewett, M., … Ailles, L. E. (2014). Cell surface profiling using high-throughput flow cytometry: A platform for biomarker discovery and analysis of cellular heterogeneity. *PLoS One*, *9*(8), e105602.

González-Carvajal, S. & Garrido-Merchán, E. C. (2005). Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012. 2020 May 26.

He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems*, *212*, 106622.

Jamin, C., Le Lann, L., Alvarez-Errico, D., Barbarroja, N., Cantaert, T., Ducreux, J., Dufour, A. M., Gerl, V., Kniesch, K., Neves, E., Trombetta, E., Alarcón-Riquelme, M., Marañon, C., & Pers, J. O. (2016). Multi-center harmonization of flow cytometers in the context of the European "PRECISESADS" project. *Autoimmunity Reviews*, *15*(11), 1038–1045.

Kadra, A., Lindauer, M., Hutter, F., & Grabocka, J. (2021). Well-tuned simple nets excel on tabular datasets. *Advances in neural information processing systems*, 34.

Larbi, A. (2017). Flow cytometry in multi-center and longitudinal studies. In *Single cell analysis* (pp. 119–132). Singapore: Springer.

Liu, H., Hu, Z. Z., Torii, M., Wu, C., & Friedman, C. (2006). Quantitative assessment of dictionary-based protein named entity tagging. *Journal of the American Medical Informatics Association*, *13*(5), 497–507.

Maecker, H. T., JP, M. C., Jr., FOCIS Human Immunophenotyping Consortium, Amos, M., Elliott, J., Gaigalas, A., Wang, L., Aranda, R., Banchereau, J., Boshoff, C., Braun, J., Korin, Y., Reed, E., Cho, J., Hafler, D., Davis, M., Fathman, C. G., Robinson, W., Denny, T., … Yeh, J. H. (2010). A model for harmonizing flow cytometry in clinical trials. *Nature Immunology*, *11*(11), 975–978.

Mäkinen, S., Skogström, H., Laaksonen, E., & Mikkonen, T. (2021). Who needs MLOps: What data scientists seek to accomplish and how can MLOps help? arXiv: 2103.08942.

Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M., & Gao, J. (2021). Deep learning-based text classification: A comprehensive review. *ACM Computing Surveys*, *54*(3), 1–40.

Montante, S., & Brinkman, R. R. (2019). Flow cytometry data analysis: Recent tools and algorithms. *International Journal of Laboratory Hematology*, *41*(Suppl 1), 56–62.

Moulard, M., & Ozoux, M. L. (2016). How validated receptor occupancy flow cytometry assays can impact decisions and support drug development. *Cytometry. Part B, Clinical Cytometry*, *90*(2), 150–158.

Natale, D. A., Arighi, C. N., Barker, W. C., Blake, J. A., Bult, C. J., Caudy, M., Drabkin, H. J., D'Eustachio, P., Evsikov, A. V., Huang, H., Nchoutmboube, J., Roberts, N. V., Smith, B., Zhang, J., & Wu, C. H. (2011). The protein ontology: A structured representation of protein forms and complexes. *Nucleic Acids Research*, *39*(Database issue), D539–D545.

Nolan, J. P., Condello, D., Duggan, E., Naivar, M., & Novo, D. (2013). Visible and near infrared fluorescence spectral flow cytometry. *Cytometry*, *83A*, 253–264.

Olson, R. S. & Moore, J. H. (2016). TPOT: A tree-based pipeline optimization tool for automating machine learning. In *Proceedings of the workshop on automatic machine learning* (Vol. 64, pp. 66–74).

Overton, J. A., Vita, R., Dunn, P., Burel, J. G., Bukhari, S. A. C., Cheung, K. H., Kleinstein, S. H., Diehl, A. D., & Peters, B. (2019). Reporting and connecting cell type names and gating definitions through ontologies. *BMC Bioinformatics*, *20*(Suppl 5), 182.

Robinson, J. P. (2019). Spectral flow cytometry—*Quo vadimus*? *Cytometry*, *95*, 823–824.

Rodriguez-Esteban, R. (2015). Biocuration with insufficient resources and fixed timelines. *Database: The Journal of Biological Databases and Curation*, *2015*, bav116.

Rodriguez-Esteban, R. (2016). Understanding human disease knowledge through text mining: What is text mining? In W. Loging (Ed.), *Bioinformatics and computational biology in drug discovery and development*. Cambridge University Press.

Rodriguez-Esteban, R. (2019). Text mining applications. In *Encyclopedia of bioinformatics and computational biology* (Vol. 3, pp. 996–1000). Elsevier.

Rodriguez-Esteban, R., & Loging, W. T. (2013). Quantifying the complexity of medical research. *Bioinformatics*, *29*(22), 2918–2924.

Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84–90.

White, S., Laske, K., Welters, M. J., Bidmon, N., van der Burg, S. H., Britten, C. M., Enzor, J., Staats, J., Weinhold, K. J., Gouttefangeas, C., &

Chan, C. (2015). Managing multi-center flow cytometry data for immune monitoring. *Cancer Informatics*, *13*(Suppl 7), 111–122.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J. W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, *3*, 160018.

How to cite this article: Rodriguez-Esteban, R., Duarte, J., Teixeira, P. C., Richard, F., Koltsova, S., & So, W. V. (2022). Prediction of standard cell types and functional markers from textual descriptions of flow cytometry gating definitions using machine learning. *Cytometry Part B: Clinical Cytometry*, 1–8. https://doi.org/10.1002/cyto.b.22065