Towards Rapid Prediction of Drug-resistant Cancer Cell Phenotypes: Single Cell Mass Spectrometry Combined with Machine Learning

Journal:	*ChemComm*
Manuscript ID	CC-COM-10-2018-008296.R2
Article Type:	Communication

Journal Name

COMMUNICATION

Received 00th January 20xx,
Accepted 00th January 20xx

DOI: 10.1039/x0xx00000x

www.rsc.org/

Towards Rapid Prediction of Drug-resistant Cancer Cell Phenotypes: Single Cell Mass Spectrometry Combined with Machine Learning

Renmeng Liu,[‡a] Genwei Zhang[‡a] and Zhibo Yang*[a]

Combined single cell mass spectrometry and machine learning methods is demonstrated for the first time to achieve rapid and reliable prediction of the phenotype of unknow single cells based on their metabolomic profiles, with experimental validation. This approach can be potentially applied towards prediction of drug-resistant phenotypes prior to chemotherapy.

Drug resistance, a phenomenon that renders tumor evasion of anticancer agents, is regarded as the major reason for chemotherapeutic failures.[1] In other words, a small population of cells capable of surviving from chemo-treatment through complex drug-resistant mechanisms, become immune to the original therapy, and eventually induce cancer relapse.[2] In general, there are two major types of drug resistance: primary and acquired. Primary resistance reduces the efficacy of chemotherapies before drug exposure, whereas acquired drug resistance develops afterwards.[3] Unfortunately, drug resistance cannot be monitored or evaluated in advance using common molecular imaging techniques, such as positron emission tomography, until accomplishing one or two chemo-treatment cycles in modern clinical practice,[4] resulting in ineffective treatment accompanied by serious toxicity for the patients. In addition, different tumor cells within the same histological region may respond differently to chemo-treatment due to intratumor heterogeneity.[5] However, conventional studies of drug resistance based on cell populations lack the ability to uncover biological information masked by such tumor cell heterogeneity. Herein, it is imperative to study drug resistance through interrogation and evaluation of individual cells using single-cell based methodologies. Mass spectrometry (MS) is a fast developing technique with broad applications in fundamental science and biomedical studies.[6] Recent development in MS allows for analysis of single cells with limited amount of analytes available (as low as in pL range for mammalian cells)[7] due to its extraordinary sensitivity, high accuracy, and high throughput. To date, reported single cell MS (SCMS) techniques include but are not limited to secondary ion MS (SIMS),[8] matrix-assisted laser desorption/ionization (MALDI) MS,[9] laser ablation electrospray ionization (LAESI) MS,[10] live-single cell video-MS,[11] induced nanoESI MS,[12] the Single-probe MS,[13] and the T-probe MS.[14] Among these techniques, the Single-probe MS method stands out as an ambient technique to analyze live single cells of interest in situ and in real time with high efficiency and reliability.[13, 15]

On the other hand, cell adhesion-mediated drug resistance (CAM-DR) was reported for myelogenous leukemia cells upon adhering to extracellular matrix (ECM), which coexists with those leukemic cells in the bone marrow, through integrin-ECM interaction.[16] Interestingly, this cell-ECM interaction confers reduced cell apoptosis upon exposure to cytotoxic drugs, and was recognized as one important form of primary drug resistance.[17] Despite the achievements of illustrating related biological mechanisms,[18] limited effort was contributed to predict such drug-resistant phenotype prior to any chemo-treatment, exposing patients to the risk of ineffective chemotherapy and associated toxicity. Limited studies in this area are likely due to a variety of factors, including 1) the lack of rapid and sensitive single cell analytical approaches that can simultaneously unveil phenotypical discrimination and intratumor heterogeneity, 2) the shortage of methods for systematic metabolomic analysis of single cells to reveal cellular metabolomic profiles associated with different phenotypes, and 3) the absence of advanced data mining methods towards rapid and reliable prediction.

To address those issues, we used the Single-probe SCMS technique to conduct metabolomic analysis at single cell level (i.e., single cell metabolomics) of cultured chronic myelogenous leukemia (CML) cells (K-562) and obtain metabolomic information that is sensitive to upstream gene expression, protein regulation, and change of surrounding

[a] *Department of Chemistry and Biochemistry, University of Oklahoma, Norman, Oklahoma, US, 73019*

‡These authors contributed to equally to this communication.

†Electronic Supplementary Information (ESI) available: [Experimental details, SCMS data analysis, ML procedures and supporting tables]. See DOI: 10.1039/x0xx00000x

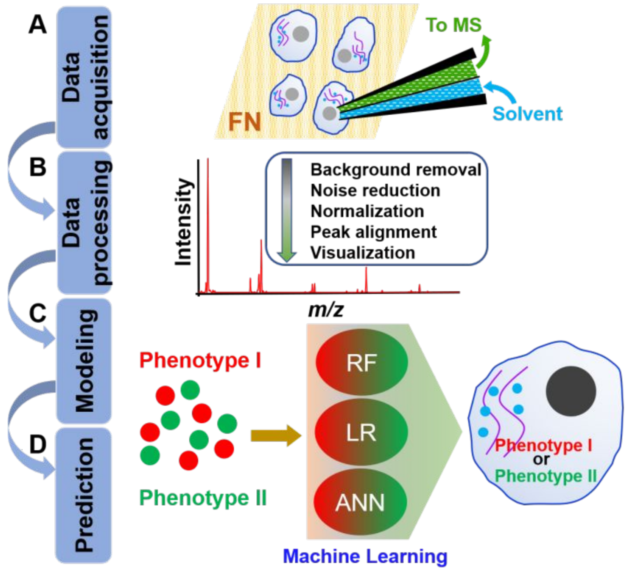


Fig. 1 Workflow of the combined single cell mass spectrometry (SCMS) experiments and machine learning (ML) data analysis methods. (A) MS measurements of single cells using the Single-probe SCMS technique. (B) A comprehensive data processing approach to extract metabolomic information from raw SCMS datasets and visualize cellular profiles in low dimensional space. (C) ML models built on cells with two different phenotypes (with or without CAM-DR). (D) Rapid and reliable prediction of drug-resistant phenotypes at single cell level.

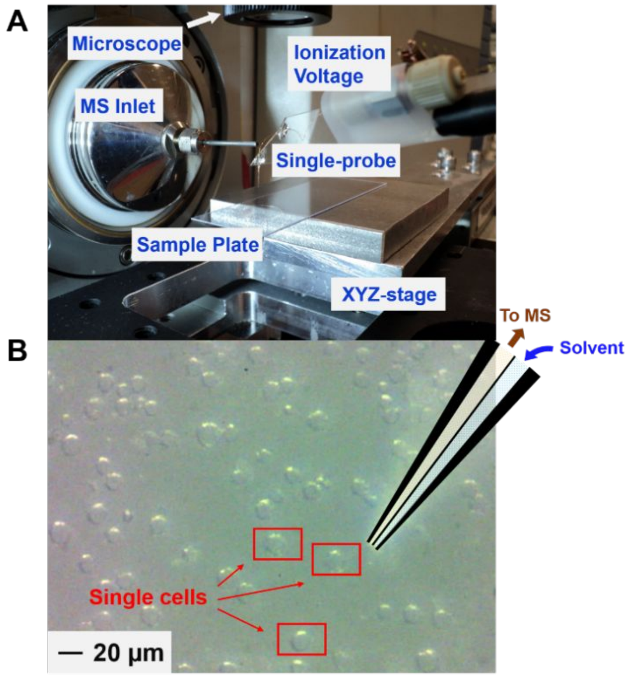


Fig. 2 (A) Experimental setup of the SCMS platform, which is an integrated system including a Single-probe device, a Thermo Orbitrap XL mass spectrometer, two microscopes, and a motorized XYZ-stage system. (B) Individual leukemic cells located on the sample plate to be analyzed.

microenvironment.[19] Data analysis was conducted using machine leaning (ML) algorithms to mine the complex metabolomic datasets and unveil hidden biological patterns by performing clustering, regression, and prediction.[20] To the best of our knowledge, it is the first time to combine SCMS experiments with ML models for single cell metabolomics studies. Our approach provides a potential solution towards rapid and reliable prediction of drug-resistant cancer cell phenotypes (e.g. CAM-DR) based on cellular metabolomic profiles.

K-562 cell line was used as a model system to demonstrate our strategy as shown in Fig. 1. As a well-established model, this cell line has been previously used to study the mechanism of CAM-DR in cancer cells.[16-18] We followed the published protocols to prepare two different phenotypes.[16] In brief, we first coated glass cover slips with fibronectin (FN), a major component of ECM,[21] and then allowed CML cells (K-562 cell line) to interact with FN in the cell culture plate. Cells that can adhere to FN (phenotype I) were reported to present CAM-DR compared with those suspended in the culture medium (phenotype II).[16] We prepared single cells of both phenotypes on the same type of glass cover slips (see “Cell Culture and Sample Preparation” in ESI†) Using a hemocytometer, we estimated that 23.9% ± 5.3% of cells possessed CAM-DR in a typical experiment. We then utilized the Single-probe SCMS platform (Fig. 2A and “SCMS Experiments” in ESI†) to interrogate individual cells and obtained their corresponding metabolomic profiles in real-time analysis (Fig. 2B). We analyzed 100 and 108 single cells of phenotypes I and II, respectively. The raw MS data were subjected to pre-treatment, including background removal, noise reduction, peak normalization, and peak alignment (as described in “SCMS Data Analysis” in ESI†). The endogenous cellular metabolites along with their relative ion intensities were subjected to downstream comprehensive analyses, including statistical analyses and ML predictions.

To qualitatively evaluate and visualize the difference of metabolomic profiles between these two phenotypes, we analyzed the SCMS data using the *t*-distributed stochastic neighbor embedding (*t*-SNE), an algorithm for dimensionality reduction and visualization of data points in a non-linear fashion to achieve subtle group discrimination.[22] As shown on the *t*-SNE plot (Fig. 3), an evident discrimination between these two phenotypes can be intuitively observed, although some overlapped data points still exist likely due to cell heterogeneity. Our results suggest that the metabolomic profiles of two phenotypes are significantly different, which might be attributed to integrin-ECM interaction. With such evident discrimination, we further applied ML algorithms to establish models capable of predicting cellular phenotypes (i.e., CAM-DR or non-CAM-DR) based on the metabolomic profiles of cells.

In our study, we constructed ML models using random forest (RF), penalized logistic regression (LR), and artificial neural network (ANN) following SCMS data pre-treatment as described earlier. RF is an ensemble learning method based on multiple constructed decision trees and eventually outputs the averaged decision. Penalized LR builds nonlinear relationship between the response variable and independent variables through a logistic function, followed by minimizing the impact of less contributing variables. Both RF and penalized LR methods have

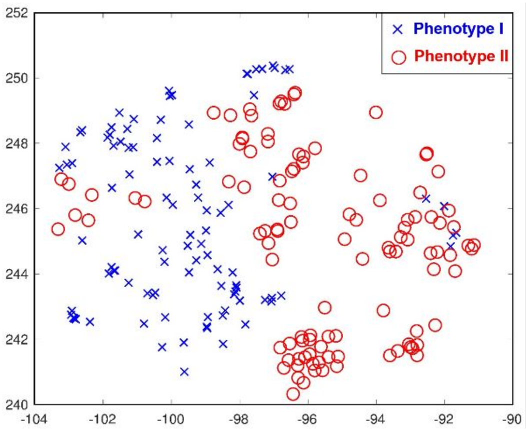


Fig. 3 Visualization of cellular metabolomic profiles in two-dimensional space using t-distributed stochastic neighbor embedding (t-SNE). Phenotypic discrimination between two types of cells (phenotype I and II) is evident.

been broadly applied to conventional metabolomic studies using liquid chromatography-mass spectrometry (LC-MS)[23] and single cell RNA-seq datasets.[24] ANN, as a fast-developing ML method, was inspired by the biological neural networks in animal brains. ANN optimizes parameters by learning from the prior knowledge, and the optimized model generates predictions through connected units and nodes. ANN has been previously applied to sorting single cells based on measured biomechanical properties,[25] and prediction of patient survival though genomics data.[26] Here, we further expanded the applications of those three methods to the analysis of single cell metabolomics datasets obtained from the Single-probe SCMS technique. Specifically, we applied RF, penalized LR (i.e., elastic net LR), and ANN to our pre-treated single cell metabolomics datasets, evaluated the predictive accuracy of each ML model, and recorded the demanded computing time under each experimental condition. We performed model construction, evaluation, and k-fold validation for each ML model (see "Machine Learning (ML) and Model Evaluation" and Tables S1–S3 in ESI†).

The pre-treated datasets were randomly shuffled with 80% cells being selected as the training set and the remaining 20% being selected as the testing set. The training set was used to construct and train ML models, whereas the testing set was used to evaluate the model performance. Due to tumor cell heterogeneity and experimental variation, single cell metabolomics datasets contain missing values (i.e., undetected cellular metabolites that were labelled as in 0 values in SCMS metabolomics datasets) in some SCMS measurements. Therefore, we evaluated the model performance according to different missing value threshold (MVT) as shown in Fig. 4. For example, a dataset with 20% MVT contains variables (metabolites) that can be detected in at least 80% of all measured single cells. As the MVT increases, the number of variables increases accordingly (i.e., from 7 to 3232 as the MVT increases from 0% to 90%) in each ML model. A gradually improved predictive accuracy was also observed in all three models (Fig. 4A–4C). Notably, a pronounced improvement was observed in predictive accuracy (from 77.1% ± 10.2% to the highest value of 94.8% ± 4.2%) in the RF model when the MVT was raised from 0% to 50%. However, further improvement of predictive accuracy was not observed with higher MVTs. Compared with the RF model, both penalized LR and ANN methods produced higher predictive accuracy when the MVT was below 40%, whereas comparable predictive accuracy was achieved as the MVT exceeded 40%. In addition, the highest predictive accuracy (i.e., 94.7% ± 1.8% and 96.2% ± 2.7%) can be obtained at 80% MVT for penalized LR and ANN models, respectively. Considering the trade-off between predictive accuracy and computing cost, which is a critical factor when handling larger sizes of data, we adopted ANN model with 40% MVT for rapid (~ 6 s) and reliable prediction (> 95% predictive accuracy) of drug-resistant phenotypes. We further demonstrated the predictive power of all ML models (with 40% MVT) in distinguishing two phenotypes using receiver operating characteristic (ROC) curve analysis[27] that examines the sensitivity and specificity of the model (Fig. 4D–4F). Consistently, the ANN model is superior in prediction with the area under the curve (AUC_{ANN}) = 0.9976 compared with the other two models (AUC_{RF} = 0.9542 and $AUC_{penalized\ LR}$ = 0.9884). To experimentally validate our method and evaluate the predictive accuracy of the ANN model, we conducted SCMS experiments and data pre-treatment for another batch of 31 single cells prepared on a different day, and utilized the trained ANN model to predict this new set of data (see "Method Validation" in ESI†). Our results show that the ANN model

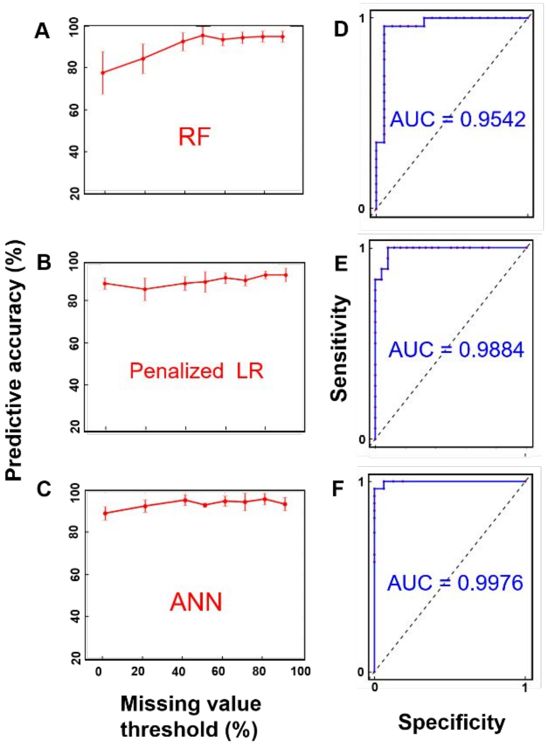


Fig. 4 Evaluation of ML models. (A–C) Predictive accuracy of the random forest (RF), penalized logistic regression (LR), and artificial neural network (ANN) model were evaluated using different missing value thresholds (MVTs). (D–F) Evaluation of the predictive power of the corresponding RF, penalized LR, and ANN models using receiver operating characteristic (ROC) curve analysis. The area under the curve (AUC) were calculated for all three models.

produced 87.1% ± 4.8% predictive accuracy, achieving a comparable performance compared with our earlier results on the testing set.

In conclusion, we reported studies using the combined ambient SCMS technique (i.e., the Single-probe MS) and ML models to distinguish and predict drug-resistant phenotypes (e.g. CAM-DR) of live single cells through cellular metabolomic profiles for the first time. Previous studies reported a number of prediction methods based on metabolic biomarkers (i.e., cellular species characteristic of specific disease, phenotype, etc.), including two-sample *t*-test,[28] analysis of variance (ANOVA),[23b] loadings of principle component analysis (PCA),[29] and orthogonal partial least squares-discriminant analysis (OPLS-DA).[14] Compared with the above reported models, our method presents the following unique advantages: 1) SCMS based experiments allow for recognition of heterogeneous cells with different phenotypes. 2) Minimum sample preparation enables metabolomic signatures of live cells to be captured through online and in situ measurements. 3) Constructed ML models provide rapid results, which facilitates their potential translational applications towards future point-of-care (POC)[30] prognostic assays. 4) Because our methods utilized a variety of cellular metabolites other than metabolic biomarkers alone, the model predictive accuracy is significantly improved (*p*-value < 0.05, from Welch's one-tail *t*-test) compared with other models utilizing biomarkers discovered through two-sample *t*-test or PCA loading plot (see "Statistical Analyses", "Model Comparison" and Fig. S1 in ESI†). As a complementary approach to biomarker identification at the population level, LC-MS/MS analysis of cell lysate was performed. Among all discovered biomarkers from single cells (e.g., 70 metabolites obtained from *t*-test), 28 of them were identified using LC-MS/MS (see "Identification of Metabolic Biomarkers" and Table S4 in ESI†). This complementary method can potentially benefit future SCMS studies, although all species of interest in single cells may not be identified from cell lysates, likely due to rapid metabolite turnover during sample preparation.[7] In addition, we validated our methods using cells prepared from different batches to obtain comparable results. Although the cultured CML cells were used as the model in the current study, our method can be potentially used towards future prediction and prognosis of patient derived samples. However, because the clinical samples are rather complex, additional procedures for sample preparation are necessary. For example, heterogenous cells obtained from bone marrow biopsy in clinic need to be firstly purified, followed by enrichment of leukemic cells using standard protocols including centrifugation and flow cytometry analysis[31] prior to the SCMS experiments (~ 30 s/cell) and ML predictions of drug-resistant phenotypes.

Conflicts of interest

The authors declare no conflicts of interests.

Notes and references

1 M. M. Gottesman, *Annu. Rev. Med.*, 2002, **53**, 615-627.

2 L. A. Garraway and P. A. Janne, *Cancer Discov.*, 2012, **2**, 214-226.

3 H. Zahreddine and K. L. B. Borden, Front Pharmacol, 2013, 4.

4 T. H. Lippert, H. J. Ruoff and M. Volm, *Int. J. Med. Sci.*, 2011, **8**, 245-253.

5 B. Zhao, J. R. Pritchard, D. A. Lauffenburger and M. T. Hemann, *Cancer Discov.*, 2014, **4**, 166.

6 a) X. J. Feng, X. Liu, Q. M. Luo and B. F. Liu, *Mass Spectrom. Rev.*, 2008, **27**, 635-660; b) F. W. McLafferty, *Annu. Rev. Anal. Chem.*, 2011, **4**, 1-22.

7 L. Zhang and A. Vertes, *Angew. Chem. Int. Ed.*, 2018, **57**, 4466-4477.

8 N. Musat, R. Foster, T. Vagner, B. Adam and M. M. M. Kuypers, *Fems Microbiol. Rev.*, 2012, **36**, 486-511.

9 A. J. Ibáñez, S. R. Fagerer, A. M. Schmidt, P. L. Urban, K. Jefimovs, P. Geiger, R. Dechant, M. Heinemann and R. Zenobi, *Proc. Natl. Acad. Sci. U. S. A.*, 2013, **110**, 8790.

10 B. Shrestha and A. Vertes, *Anal. Chem.*, 2009, **81**, 8265-8271.

11 T. Masujima, *Anal. Sci.*, 2009, **25**, 953-960.

12 H. Zhu, G. Zou, N. Wang, M. Zhuang, W. Xiong and G. Huang, *Proc. Natl. Acad. Sci. U. S. A.*, 2017, **114**, 2586.

13 N. Pan, W. Rao, N. R. Kothapalli, R. M. Liu, A. W. G. Burgett and Z. B. Yang, *Anal. Chem.*, 2014, **86**, 9376-9380.

14 R. Liu, N. Pan, Y. Zhu and Z. Yang, *Anal. Chem.*, 2018, **90**, 11078-11085.

15 a) N. Pan, W. Rao, S. J. Standke and Z. B. Yang, *Anal. Chem.*, 2016, **88**, 6812-6819; b) W. Rao, N. Pan and Z. Yang, *J. Vis. Exp.*, 2016, **112**, 53911.

16 J. S. Damiano, L. A. Hazlehurst and W. S. Dalton, *Leukemia*, 2001, **15**, 1232.

17 K. H. Shain and W. S. Dalton, *Mol. Cancer Ther.*, 2001, **1**, 69.

18 a) J. S. Damiano, A. E. Cress, L. A. Hazlehurst, A. A. Shtil and W. S. Dalton, *Blood*, 1999, **93**, 1658-1667; b) L. A. Hazlehurst and W. S. Dalton, *Cancer Metast. Rev.*, 2001, **20**, 43-50.

19 G. J. Patti, O. Yanes and G. Siuzdak, *Nat. Rev. Mol. Cell Biol.*, 2012, **13**, 263.

20 D. Grapov, J. Fahrmann, K. Wanichthanarak and S. Khoomrung, *OMICS*, 2018, **22**, 630.

21 P. A. Harper, P. Brown and R. L. Juliano, *J. Cell Sci.*, 1983, **63**, 287.

22 a) T. D. Do, T. J. Comi, S. J. B. Dunham, S. S. Rubakhin and J. V. Sweedler, *Anal. Chem.*, 2017, **89**, 3078-3086; b) X. Li, W. Chen, Y. Chen, X. Zhang, J. Gu and M. Q. Zhang, *Nucleic Acids Res.*, 2017, **45**, e166.

23 a) B. Xi, H. Gu, H. Baniasadi and D. Raftery, *Methods Mol. Biol.*, 2014, **1198**, 333-353; b) D. Grissa, M. Pétéra, M. Brandolini, A. Napoli, B. Comte and E. Pujos-Guillot, *Front. Mol. Biosci.*, 2016, **3**, 30.

24 M. B. Pouyan and D. Kostka, *Bioinformatics*, 2018, **34**, i79-i88.

25 E. M. Darling and F. Guilak, *Tissue Eng. Pt. A*, 2008, **14**, 1507-1515.

26 T. Ching, X. Zhu and L. X. Garmire, *PLOS Comput. Biol.*, 2018, **14**, e1006076.

27 J. G. Xia, D. I. Broadhurst, M. Wilson and D. S. Wishart, *Metabolomics*, 2013, **9**, 280-299.

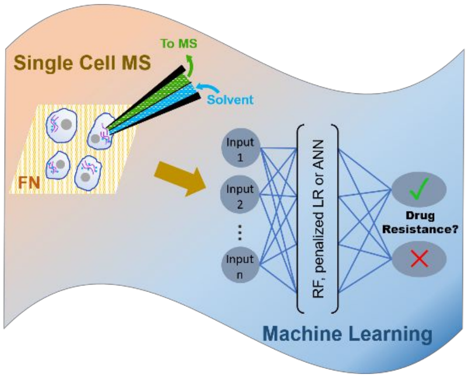
28 D. J. Hinton, M. S. Vázquez, J. R. Geske, M. J. Hitschfeld, A. M. C. Ho, V. M. Karpyak, J. M. Biernacka and D.-S. Choi, *Sci. Rep.*, 2017, **7**, 2496.

29 P. Nemes, A. M. Knolhoff, S. S. Rubakhin and J. V. Sweedler, *Anal. Chem.*, 2011, **83**, 6810-6817.

30 C. R. Ferreira, K. E. Yannell, A. K. Jarmusch, V. Pirro, Z. Ouyang and R. G. Cooks, *Clin. Chem.*, 2016, **62**, 99.

31 J. Cloos, J. R. Harris, J. J. W. M. Janssen, A. Kelder, F. Huang, G. Sijm, M. Vonk, A. N. Snel, J. R. Scheick, W. J. Scholten, J. Carbaat-Ham, D. Veldhuizen, D. Hanekamp, Y. J. M. Oussoren-Brockhoff, G. J. L. Kaspers, G. J. Schuurhuis, A. K. Sasser and G. Ossenkoppele, *J. Vis. Exp.*, 2018, **133**, e56386.

The combination of single cell mass spectrometry with machine learning enables prediction of drug-resistant cell phenotypes based on metabolomic profiles.


Single Cell MS
To MS
Solvent
FN
Input 1
Input 2
⋮
Input n
RF, penalized LR or ANN
✓
Drug Resistance?
✗
Machine Learning