Machine learning for gene expression-based prediction of individual drug response for cancer patients

Nicolas Borisov [1,*], Victor Tkachev [2,3], Maxim Sorokin [2,3], and Anton Buzdin [2,3,4]

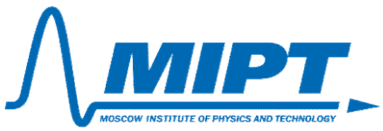
[1]Moscow Institute of Physics and Technology, 141701 Moscow Oblast, Russia

[2]OmicsWayCorp, 91788 Walnut, CA, USA,

[3]I.M. Sechenov First Moscow State Medical University, 119991 Moscow, Russia

[4]Shemyakin-Ovchinnikov Institute of Bioorganic Chemistry, 117997 Moscow, Russia

* Corresponding author: borisov@oncobox.com



Abstract: (1) Background: Various machine learning (ML) methods are applied for prediction of individual clinical efficiency of cancer drugs and therapeutic regimens. (2) Methods: We proposed a next-generation ML approach termed FloWPS (FLOating-Window Projective Separator) that uses pre-processing/trimming/filtration of multi-omics features when building the ML models, in order to preclude extrapolation in the feature space. (3) Results: Using Gene Expression Omnibus (GEO), The Cancer Genome Archive (TCGA), and Tumor Alterations Relevant for GEnomics-driven Therapy (TARGET) project databases we selected 27 gene expression datasets for cancer patients, annotated with clinical response status. Using the blind/agnostic LOO approach for data trimming, we demonstrated essential improvement of ML quality metrics (AUC, sensitivity and specificity) for FloWPS-based clinical response classifiers for all global ML methods applied, such as support vector machines (SVM), random forest (RF), binomial naïve Bayes (BNB), adaptive boosting (ADA), as well as multi-level perceptron (MLP). Namely, the AUC for the treatment response classifiers increased from 0.61–0.88 range to 0.70–0.97. (4) Conclusion: Considering our ML trial with 27 clinically annotated cancer gene expression datasets, the BNB method showed best performance for data trimming and was the most effective for classifying the clinical response using multi-omics features, with minimal, median and maximal AUC values equal to 0.77, 0.86 and 0.97, respectively

Keywords: bioinformatics; personalized medicine; oncology; chemotherapy; machine learning; omics profiling.

Machine learning methods in personalized medicine

- How to classify a new patient as responder or non-responder?
- Various omics data may be used:
  - gene expression
  - mutations
  - pathway activation
  - etc.
- Machine learning has been successful in many areas: physics, banking, defense, agriculture, etc.
- Yet, still no robust classifier in personalized oncology.

ECB 2021

Machine learning in personalized medicine…

… often fails because of:

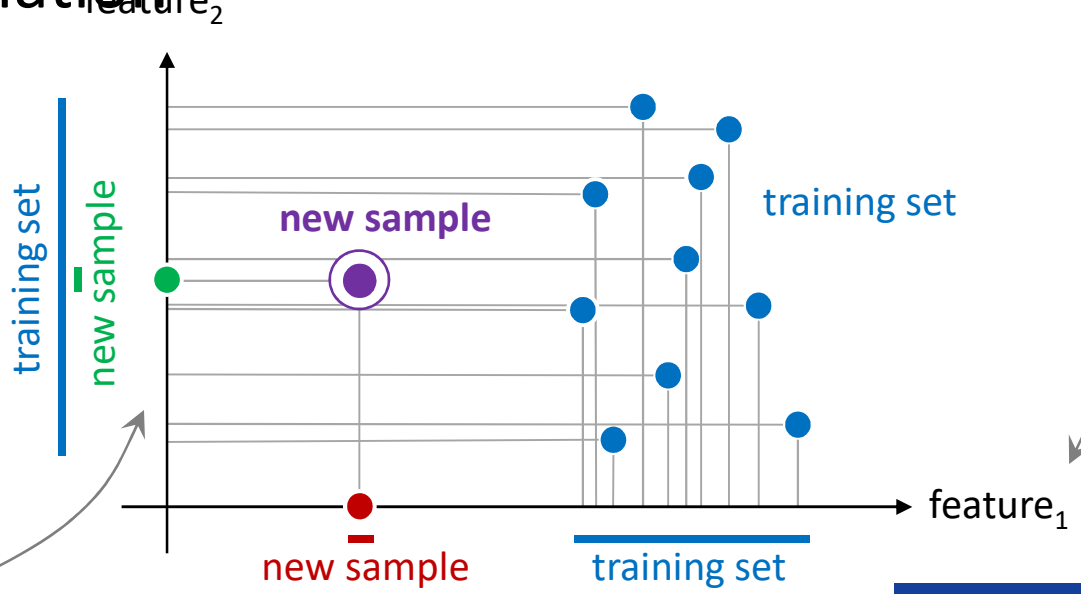
- We developed a robust approach to machine learning in personalized medicine, termed Flexible Data Trimming (FDT).
- FDT avoids extrapolation by filtering irrelevant features.

ECB 2021

FDT rationale: filtering irrelevant features

- Feature selection to avoid extrapolation

Omics features:
gene expression levels,
mutation frequencies, etc.



Projection: the new sample is inside the training set
⇒ feature is relevant and included

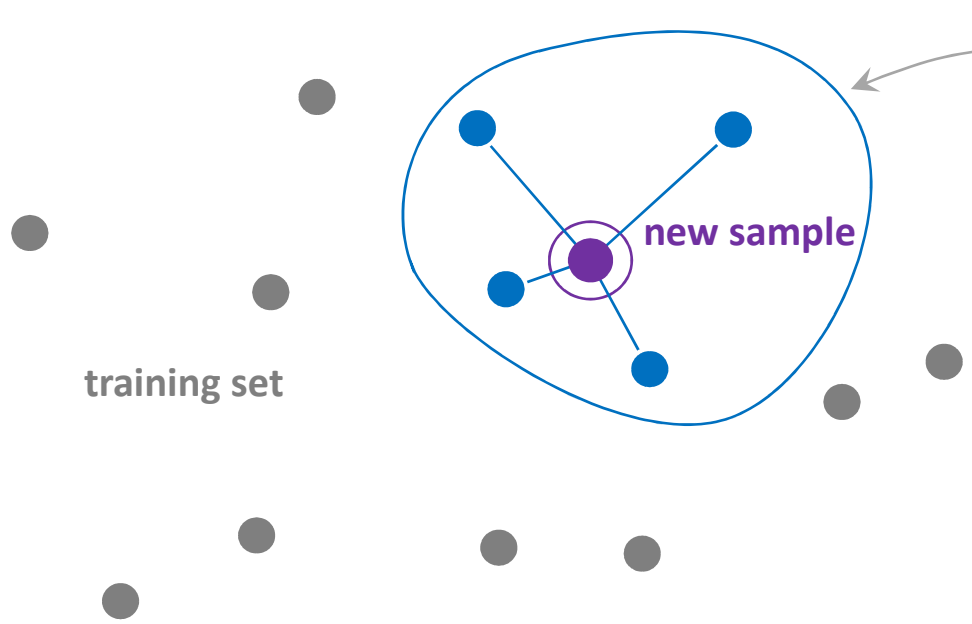
At least m points are upper and lower than the new sample projection

Projection: the new sample is outside of the training set
⇒ feature is irrelevant and not included

ECB 2021

FDT rationale: neighbors selection

To construct a machine learning model we use only k nearest training points



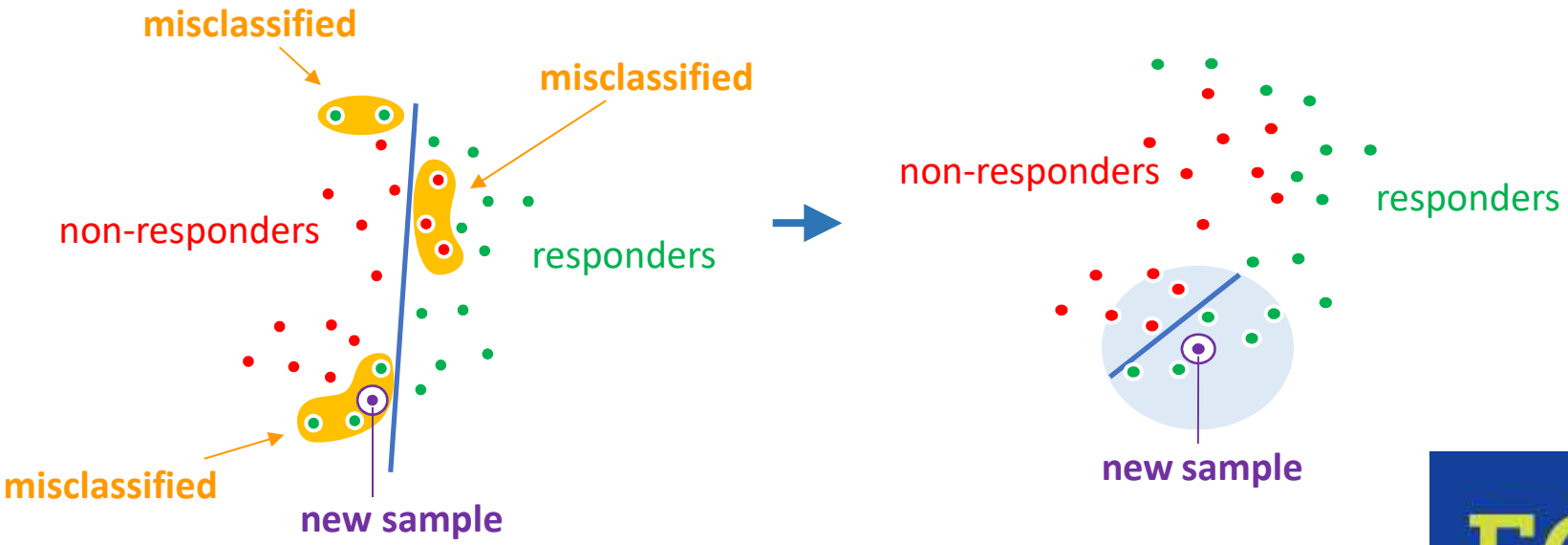
Data used for model fitting

ECB 2021

FDT rationale: a hybrid, global + local approach

- **Global machine learning methods** may fail to separate classes for datasets with no global order

Machine-learning with FDT works locally and handles that cases correctly



Machine learning with FDT is beneficial exclusively for global ML methods.

ECB 2021

Evaluation of FDT: datasets

- FDT potential has been evaluated for personalized oncology application for:
  - 2192 patients,
  - 27 treatment regimens
  - from 19 GEO, 4 TARGET, 2 TCGA datasets and 2 our own datasets

- Disease types included breast cancer (10 datasets), multiple myeloma (10 datasets), AML (3 datasets), ALL (1 dataset), Wilms kidney tumor (1 dataset), low-grade glioma (1 dataset) and lung cancer (1 dataset).

- Chemotherapeutics included taxanes, bortezomib, vincristine, trastuzumab, letrozole, tipifarnib, temozolomide, busulfan and cyclophosphamide.

ECB 2021

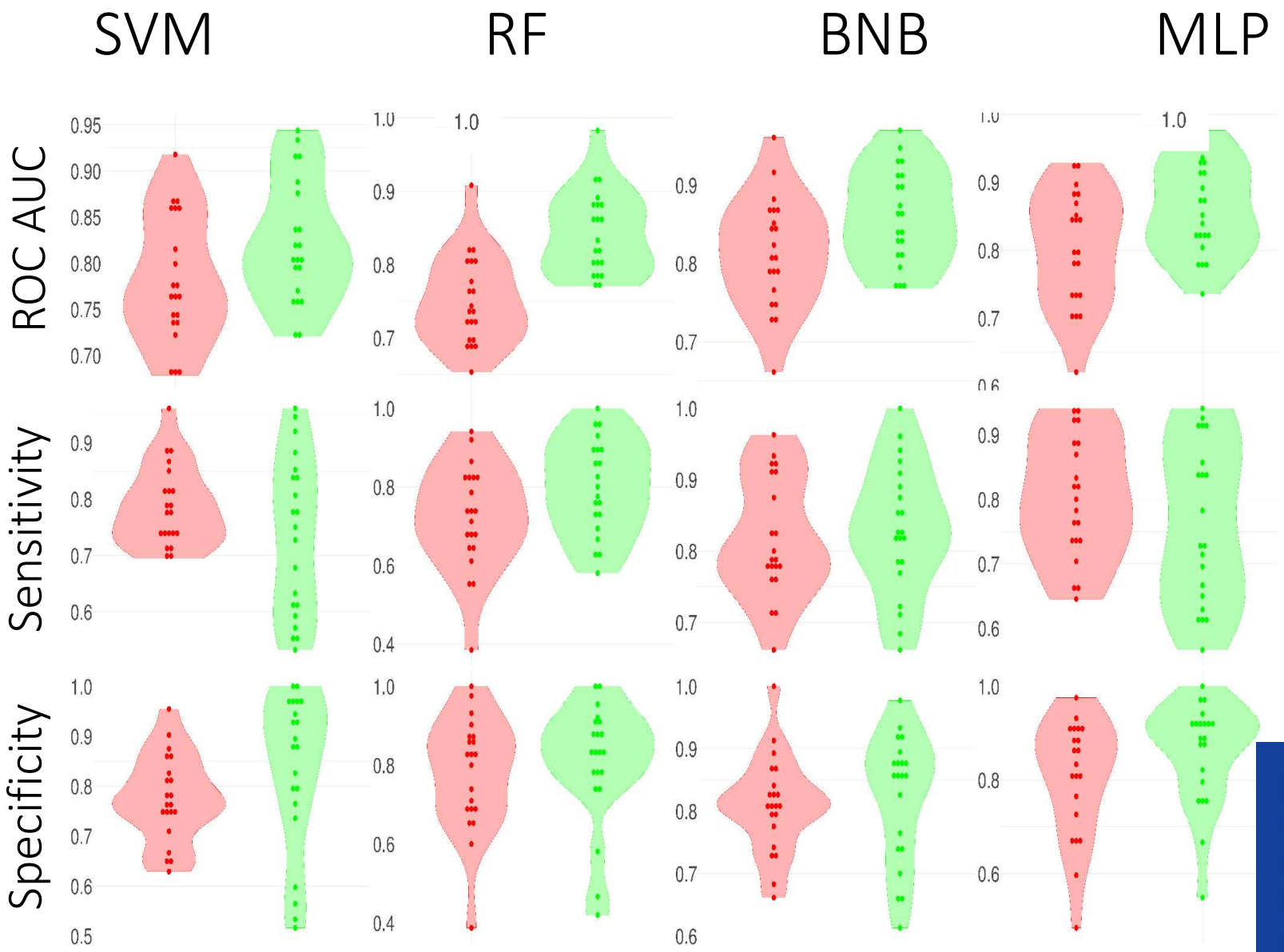
Evaluation of FDT : ML methods

- Local ML methods:
  - kNN,
  - ridge regression (RR)

- Global ML methods:
  - support vector machines (SVM),
  - random forest (RF),
  - binomial naïve bias (BNB),
  - multi-layer perceptrons (MLP),
  - adaptive boosting (ADA)

ECB 2021

Evaluation of FDT: best global ML methods



ECB 2021

Evaluation of FDT: results

- For local ML methods:
  - **kNN**,
  - **ridge regression (RR)**
- there was no advantage of FDT.

- Contrary, for global ML methods:
  - **support vector machines (SVM)**,
  - **random forest (RF)**,
  - **binomial naïve bias (BNB)**,
  - **multi-layer perceptrons (MLP)**,
  - **adaptive boosting (ADA)**
- the advantage of FDT was manifested.

- The best performance was shown by the **BNB** method.

ECB 2021

Publications

- Borisov N. et al. **Machine Learning Applicability for Classification of PAD/VCD Chemotherapy Response Using 53 Multiple Myeloma RNA Sequencing Profiles.** 2021, Front Oncol, 11:652063.doi: doi.org/10.3389/fonc.2021.652063.
- Borisov N. et al. **Cancer gene expression profiles associated with clinical outcomes to chemotherapy treatments**, 2020, BMC Medical Genomics, 13:111, doi:10.1186/s12920-020-00759-0.
- Tkachev V. et al. **Flexible Data Trimming Improves Performance of Global Machine Learning Methods in Omics-Based Personalized Oncology**. 2020, Int J Mol Sci, 21:713. doi: 10.3390/ijms21030713
- Borisov N., and Buzdin A. **New Paradigm of Machine Learning (ML) in Personalized Oncology: Data Trimming for Squeezing More Biomarkers From Clinical Datasets**. 2019, Front. Oncol. 9: 658. doi: 10.3389/fonc.2019.00658
- Tkachev V et al. **FLOating-Window Projective Separator (FloWPS): A Data Trimming Tool for Support Vector Machines (SVM) to Improve Robustness of the Classifier**, 2019, Front. Genet. 9:717. doi: 10.3389/fgene.2018.00717.

ECB 2021

Acknowledgements

- The study was supported by **Russian Scientific Foundation** Grant 21-74-20066.

- This work was supported by **Amazon** and **Microsoft Azure** grants for cloud-based computational facilities.

- We thank **Oncobox/OmicsWay** research program in machine learning and digital oncology for software and pathway databases for this study.

ECB 2021

The Team

Nicolas M. Borisov, Prof., Ph.D.
Computer science: concept development

Anton A. Buzdin, Prof., Ph.D.
Project supervision

Victor S. Tkachev
Algorithms and software

Maxim I. Sorokin, Ph.D
Bioinformatics & molecular biology

Contact to: borisov@oncobox.com

ECB 2021