Personalised drug recommendation from augmented gene expression data - the right drug(s) for the right patient

Manuela Salvucci, Davy Suvee, Deniz Pirincci, Dimitrius Raphael, Eryk Kropiwnicki, Giovanni Dall'olio, James O'Reilly, Katie Sanford, Marika Catapano, Marta Sarrico, Xenia Galkina, Francesca Mulas

Personalised medicine aims to match the right drug(s) to the right patient. However, this challenge is largely unsolved. Several research groups have focused on different aspects of the challenge, ranging from generating drug screening/perturbation datasets [1–4] and deriving clinically-relevant insights [1–3] to testing a variety of ML approaches [5–8]. While a large fraction of the literature has used gene expression or omics data as input to the ML models, more recently, other approaches leveraging a combination of gene expression and image features extracted from microscopy experiments have been applied demonstrating that the data types provide complementary information [9].

In this study, we propose to incorporate supervised learning sets for drug response in cell lines, animal models and human patients to provide ranked drug option recommendations on a patient-by-patient basis by integrating multiple and complementary public data sources. Ultimately, we aim to build a model that can take as input multiple types of sample-specific and drug/compound-specific features derived from data collected from multiple experimental models (cell lines, organoids, animal models and human patients) and learn drug response. Sample-specific input features include molecular data and metadata. The molecular data include genetics, transcriptomic, proteomic, microscopy and drug perturbation signatures. The metadata include disease type, histological features and cell type composition. Drug/compound-specific input features include drug doses and chemical properties such as ligandability, protein composition and folding. We aim to predict drug response which could be measured in the experimental models as either mechanistic and/or phenotypic readouts. Mechanistic readouts could include commonly used assays such as apoptosis or proliferation. Phenotypic readouts could include cell viability in cell lines, tumour growth in mouse models and clinical outcome and/or side-effects in human patients.

We implemented a supervised machine learning (ML) model that predicts whether a cell line may or may not be responsive to one particular drug (Gemcitabine) based on gene expression patterns using publicly available data from the Genomics of Drug Sensitivity in Cancer (GDSC) database. We benchmarked the application of different combinations of data pre-processing (e.g. variance filtering and latent variable models), feature engineering and multiple types of ML models architectures (e.g. linear regression, random forest, gradient boosting and neural networks). We assessed model performance using the negated mean squared error comparing the ML predicted drug-response with ground-truth measured drug-response (z-scored IC50). We found similar performance scores as those reported in the literature. We selected the model with best cross-validated performance (random forest on the top 50th percentile most variant genes) for downstream analysis. Additionally, we confirmed that our ML model was able to predict drug-response from gene expression for new unseen samples in an external validation cell line dataset (collected under different experimental conditions) and, to a lesser extent, in a patient-derived xenograft dataset (which had a different readout). Next, we applied the SHapley Additive exPlanations method (SHAP [10]) to score features importance and thus identify the most salient genes in driving the model's predictions. We assessed the biological relevance of the top ranking genes qualitatively and quantitatively. The top ranking genes were reviewed by an expert for relevance and specificity. Furthermore, we used our proprietary masked language model trained to predict targets in sentences from our corpus[11], using *[x] is related to [drug_name] response* as a query to quantitatively score the top ranking genes. Both qualitative and quantitative evaluation confirmed that top genes driving the model predictions were biologically relevant. Indeed, SLFN11, the highest ranking contributor gene, ranked 22nd in our masked language model and is an established predictive biomarker for sensitivity to drugs targeting the DNA damage response and DNA-damaging chemotherapies, such as Gemcitabine [12].

Ongoing work is focussed on improving the ML prediction model by further curating the input data and test alternative omics modalities (genomics, proteomics, metabolomics and sample clinico-pathological metadata) to be used as input either on their own or combined. Furthermore, we aim to do more extensive testing of pre-processing and features engineering as well as model architectures and to extend the model to predict drug-response, as monotherapy, to all available compounds the datasets we have QC-ed covered. Subsequently, we would extend the model to predict drug-response to drug combinations rather than monotherapy.

References

1. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* 483, 603–307 (2012).
2. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* 166, 740–754 (2016).
3. Gao, H. *et al.* High-throughput screening using patient-derived tumor xenografts to predict clinical trial drug response. *Nat. Med.* 21, 1318–1325 (2015).
4. Holbeck, S. L. *et al.* The National Cancer Institute ALMANAC: A Comprehensive Screening Resource for the Detection of Anticancer Drug Pairs with Enhanced Therapeutic Activity. *Cancer Res.* 77, 3564–3576 (2017).
5. Douglass, E. F. J. *et al.* A community challenge for a pancancer drug mechanism of action inference from perturbational profile data. *Cell Rep. Med.* 3, 100492 (2022).
6. Kuenzi, B. M. *et al.* Predicting Drug Response and Synergy Using a Deep Learning Model of Human Cancer Cells. *Cancer Cell* 38, 672-684.e6 (2020).
7. Kuru, H. I., Tastan, O. & Cicek, A. E. MatchMaker: A Deep Learning Framework for Drug Synergy Prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* 19, 2334–2344 (2022).
8. Adam, G. *et al.* Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ Precis. Oncol.* 4, 19 (2020).
9. Way, G. P. *et al.* Morphology and gene expression profiling provide complementary information for mapping cell state. *bioRxiv* 2021.10.21.465335 (2021) doi:10.1101/2021.10.21.465335.
10. Lundberg, S. M. *et al.* From Local Explanations to Global Understanding with Explainable AI for Trees. *Nat. Mach. Intell.* 2, 56–67 (2020).
11. Brayne, A., Wiatrak, M. & Corneil, D. On Masked Language Models for Contextual Link Prediction. in *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures* 87–99 (Association for Computational Linguistics, 2022). doi:10.18653/v1/2022.deelio-1.9.
12. Coleman, N., Zhang, B., Byers, L. A. & Yap, T. A. The role of Schlafen 11 (SLFN11) as a predictive biomarker for targeting the DNA damage response. *Br. J. Cancer* 124, 857–859 (2021).