D8.1

Data-driven model for molecular targets and drug repositioning

Project number	826121
Project acronym	iPC
Project title	individualizedPaediatricCure: Cloud-based virtual-patient models for precision paediatric oncology
Start date of the project	1st January, 2019
Duration	53 months
Programme	H2020-SC1-DTH-2018-1

Deliverable type	Report
Deliverable reference number	SC1-DTH-07-826121 / D8.1 / 1.0
Work package contributing to the deliverable	WP8
Due date	May 2021 – M29
Actual submission date	28th May, 2021

Responsible organisation	UNINA
Editor	Diego di Bernardo, Mario Failli
Dissemination level	PU
Revision	1.0

Abstract	We focused on the development of patient-specific computational models to predict response to small molecules and to identify drugs with potential therapeutic effects. A software was implemented in R and integrated into the iPCCP.
Keywords	drug sensitivity predictions, expression profiles, GSEA, geometric mean

The project iPC has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 826121.

Editor

Diego di Bernardo (UNINA)

Mario Failli (UNINA)

Contributors (ordered according to beneficiary numbers)

Gennaro Gambardella (UNINA)

Giuliano De Carluccio (UNINA)

Iacopo Ruolo (UNINA)

Carolina Armengol (IGTP)

Roland Kappler (LMU)

Laura Rodríguez Navas (BSC)

Executive Summary

D8.1 provides a detailed overview of the computational tool proposed for predicting patient-specific drugs with potential therapeutic benefits in paediatric cancer treatment.

D8.1 provides supporting evidence of the model goodness in predicting such patient-specific drugs.

D.8.1 provides guidelines on how to make better use of the software (principal aim of the deliverable).

D8.1 touches upon a case study being conducted in collaboration with IGTP and LMU partners aiming to identify and prioritize drugs for hepatoblastoma (HB) patients.

Table of Content

List of Figures

List of Tables

Chapter 1 Introduction

The aim of the deliverable is to report on the developed software (implemented in R and integrated into the iPCCP) that can generate a list of small molecules with potential therapeutic benefits for paediatric cancer treatment.

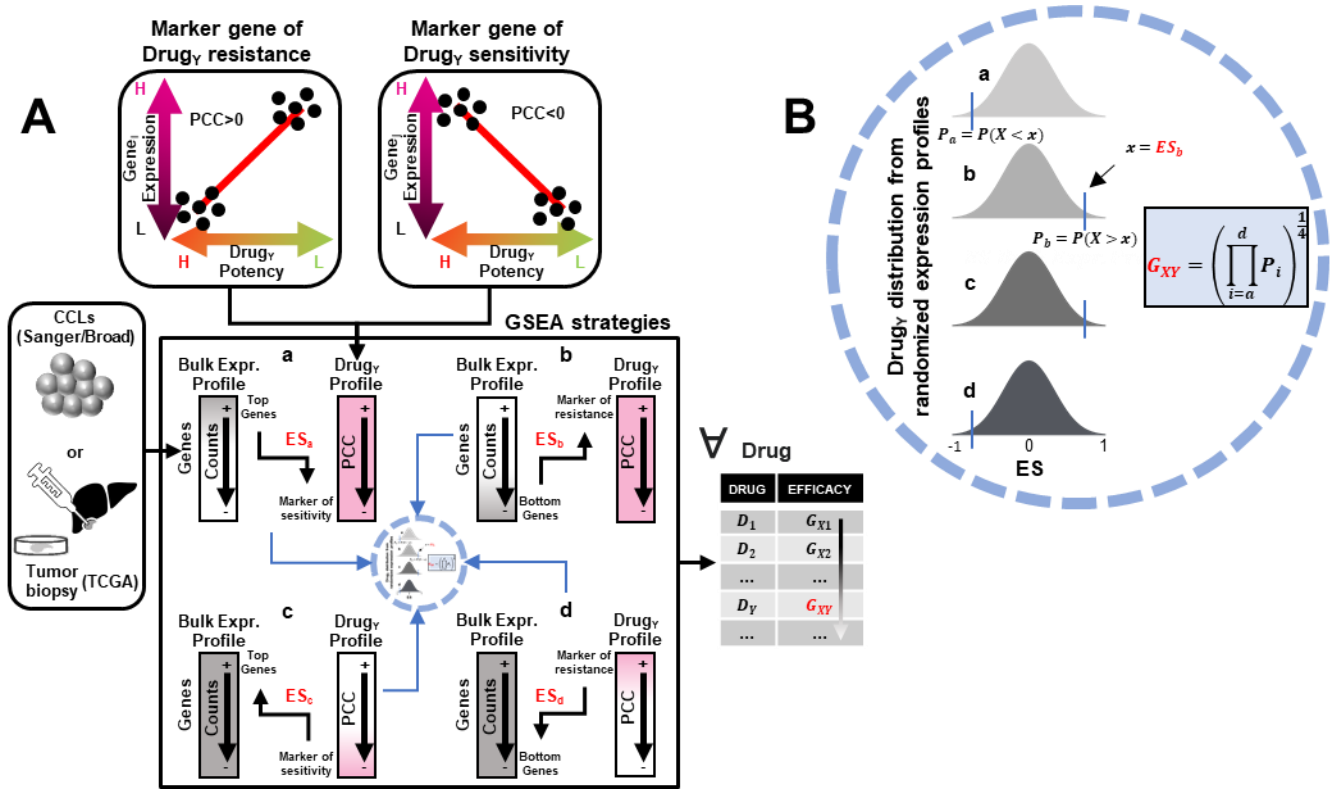

*Figure 1: Proposed method workflow. **A**) Construction of the ranked list of biomarkers. Briefly, for each gene and for each drug, the expression of the gene is correlated with the potency of the drug across the common cell-lines. Then, the enrichment of the top and bottom 250 expressed genes of each cell against the ranked list of biomarkers for each model drug (strategies a-b), and the enrichment of the best 250 marker of sensitivity and resistance of each drug against the expression profile of each cell (strategies c-d), is evaluated. **B**) ESs are compared with those resulting from randomized expression profiles. A one-tailed test (either left-tailed for strategies a and d or right-tailed for strategies b and c) is performed to assign a probability value (p-value) to each ESs. Finally, the "parallel" geometric means of p-values across the different strategies is computed to assess the overall efficacy of each drug in each cell.*

The proposed method combines bulk gene expression data and drug response data measured in cancer cell lines (CCLs) for predicting small molecule inhibitors of cancer cell growth. In details, for each gene and for each drug, the correlation between the expression of the gene and the in-vitro response to the drug across the common CCLs is computed. Drug responses, expressed either in terms of Area Under the Curve (AUC) or half maximal inhibitory concentration (IC50), reflect the potency of a drug in inhibiting the growth of a specific cancer cell line, with lower and higher values corresponding to cell sensitivity and resistance to the drug, respectively. Hence, the expression levels of putative marker genes of drug resistance in certain CCLs are positively correlated with drug response values measured in those cells (the more expressed the gene, the higher the drug concentration needed to inhibit cell growth, and the opposite); vice versa, a negative correlation denotes a putative marker gene of drug sensitivity (Figure 1A). By ranking all the genes according

to their correlation coefficient, a drug profile is produced and hereinafter used to predict drug efficacy at both cell and patient level. Specifically, to estimate the anticancer effect of that drug on a specific cancer cell line whose bulk expression profile has been generated, the enrichment extent of top/bottom expressed genes in markers of sensitivity/resistance, and vice versa, is quantified using Gene Set Enrichment Analysis (GSEA) [1] (Figure 1A, strategies a-d). Then, for each GSEA strategy, the probability of observing smaller/greater enrichment scores (ESs) from randomized expression profiles is determined independently. To note that probability directionality depends on the expected value, namely negative ESs for strategies a and d, positive for b and c. Last, the geometric mean of individual probabilities, G, is computed to assess the overall drug efficacy (Figure 1B). The smaller the G- value, the higher the drug efficacy, thereby the likelihood that the cell line is sensitive to the drug. Finally, a drug ranked list (from the most to the least effective drug) is derived by applying the same pipeline to multiple drug profiles (Figure 1A).

Chapter 2 Method validation

Baseline (i.e., before drug treatment) gene expression data and in-vitro drug response data in CCLs, released by both the Broad and Sanger Institutes [2-5], were independently employed to train complementary sets of drug profiles following the described methodology. This complementarity mostly arose from a different selection of drugs (both targeted and chemotherapeutic compounds) being screened in cell proliferation assays at the two Institutes. Specifically, ~44% of the CCLs overall profiled (560 out of 1.283 CCLs) and ~8% of the drugs totally screened (61 out of 770 drugs) were shared by the analogous datasets (Figure 2).

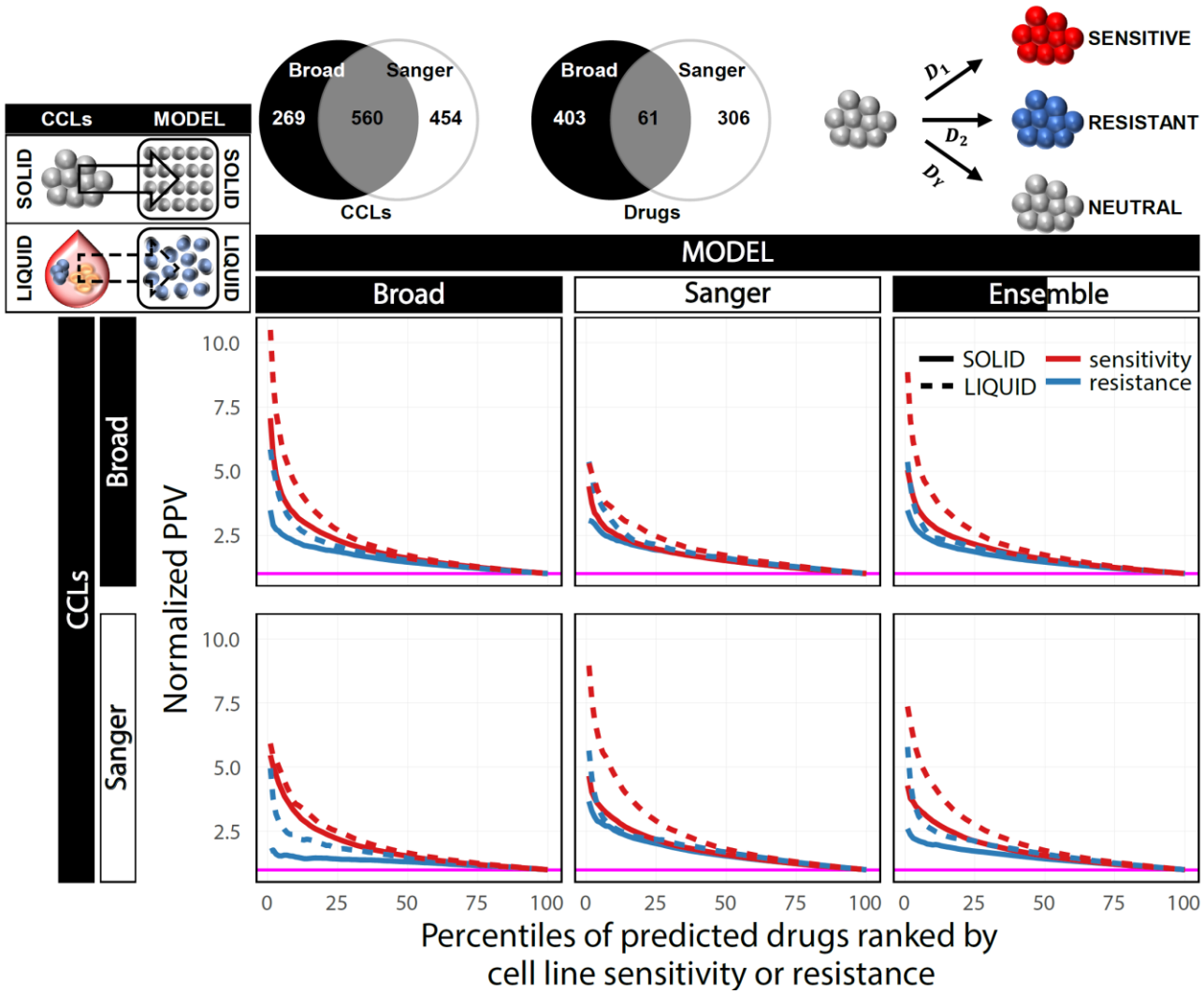

Figure 2: Validation using the proper model. From top-left to bottom-right corners: Venn Diagrams indicating overlap of CCLs and drugs between the Broad and Sanger Institutes; types of CCL responses to drug treatment; PPV curves for self, non-self and mixed CCLs-model combinations.

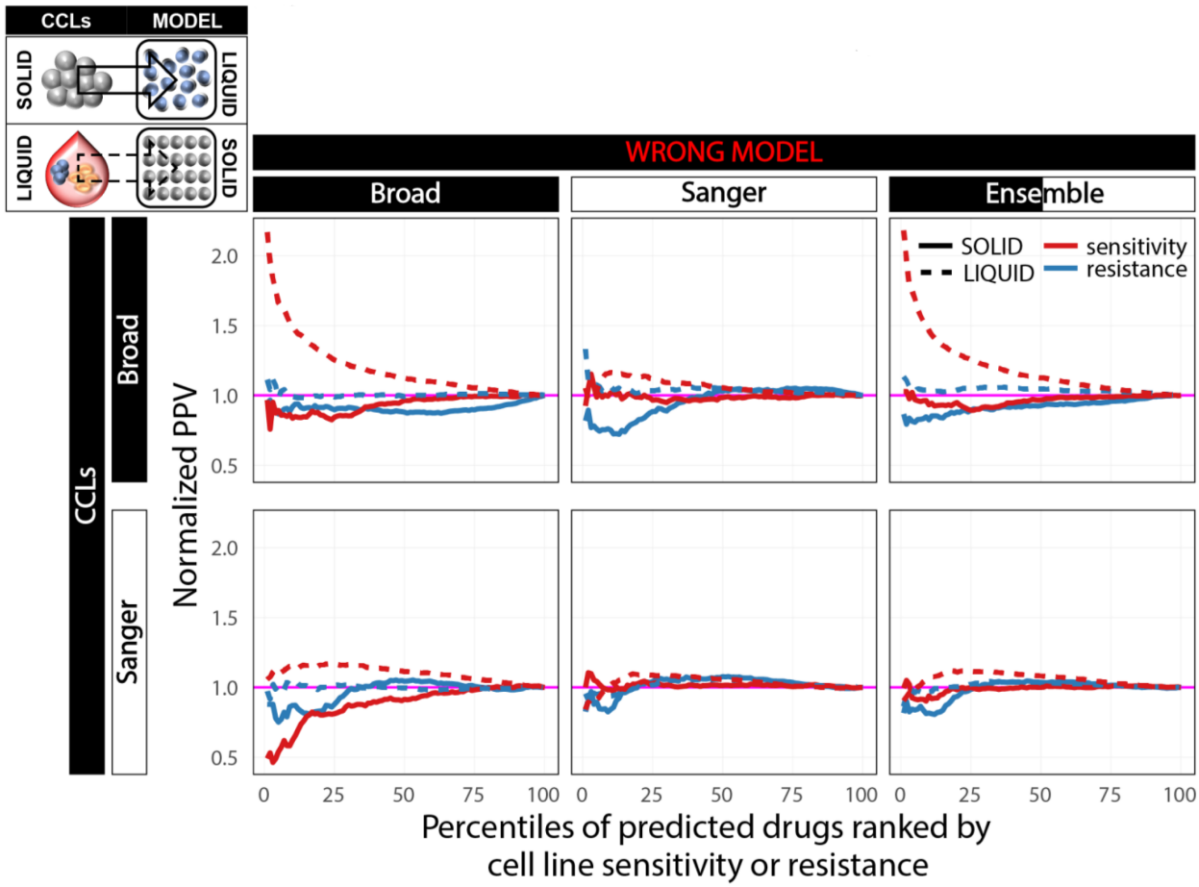

Figure 3: Validation using the wrong model. PPV curves for self, non-self and mixed CCLs-model combinations.

For each pair of Broad and Sanger data, two sets of drug profiles were derived considering CCLs from solid and liquid tumours, separately. This choice was driven by evidence of selective drug sensitivity for either solid or liquid CCLs attributable in part to differential gene expression patterns [5]. For simplicity, the sets of trained drug profiles are referred to below as solid and liquid models. To assess the predictive ability of these models, the sensitivity of solid CCLs (676 from Broad, 837 from Sanger, 461 in common) to the drugs of the Broad and Sanger solid models was estimated. The same was repeated for liquid CCLs (153 from Broad, 177 from Sanger, 99 in common) using liquid models (Figure 2). Each cell line was classified as sensitive, resistant, or neutral to a specific compound depending on drug response data, previously employed to generate the models (Figure 2). The benchmark was carried out ordering cell line-drug pair according to decreasing or increasing predicted drug efficacy. Then, the precision value (or Positive Predicted Value, PPV), namely the rate of correctly predicted drugs according to cell line sensitivity or resistance, respectively, was determined for the first n pairs, with n spanning from the 1st to the 100th percentile of total cell line-drug pairs. In Figure 2 are shown the PPV curves, normalized against the expected PPV for a random ordering of cell line-drug pairs, for self (Broad-Broad, Sanger-Sanger), non-self (Broad-Sanger, Sanger-Broad) and mixed (Broad/Sanger-Ensemble) CCLs-model combinations. Unsurprisingly, the best performances are obtained in self combinations with scores up to 10-fold better than random. This is a direct consequence of the “overfit”, since gene expression data used to train the models were also used to make predictions. Concerning non-self combinations, a significant decrease in drug resistance precision for solid CCLs is observed with the Broad rather than the Sanger model. The performances of mixed combinations, instead, are halfway between self and non-self, as expected. As a rule, sensitivity is predicted better than resistance and scores for liquid CCLs tend to be higher than solid ones. Finally, a negative control was performed by crossing solid CCLs with liquid models and the opposite. Results, shown in Figure 3, reveal a random-like trend of estimated precisions for each CCLs-model combination following drug efficacy prediction with the wrong model.

Chapter 3 Software guidelines

The software is freely available for academic use at https://vre.ipc-project.bsc.es/ under the name of DpFrEP (Drug-prediction From Expression Profiles). The following section provides guidelines on how to make better use of DpFrEP.

DpFrEP takes in input three parameters, an input file that needs to be uploaded and two additional arguments selected from specific drop-down menus. Detailed information on these parameters is provided below.

The input file contains normalized expression values for each gene in each sample. The user must ensure that the file adheres to the CSV format, that genes and samples are placed on rows and columns, respectively, and that both gene and sample identifiers are provided (Figure 4). Note that the current version of DpFrEP supports only Ensembl gene identifiers. To upload the file a registered account is required.

```
,HB11,HB28,HB33,HB48
ENSG00000164266,11.3,13.2,11.1,12.3
ENSG00000117594,10.1,7.4,6.3,7.1
ENSG00000113905,6.2,12.3,11.8,8.8
ENSG00000166741,9.5,9.6,10.9,7.2
ENSG00000139219,10.3,10.7,7.9,10.1
ENSG00000138115,5.1,7.1,5.9,6.6
ENSG00000025423,4.2,8.7,5.6,5.9
ENSG00000151012,10.4,6.8,5.3,7.7
```

Figure 4: DpFrEP input file format.

To get the best performance out of the algorithm, drug predictions for solid and liquid tumours are run separately. In this regard, it is recommended to pre-process data from solid and liquid tumours independently and to save the relative normalized expression profiles in separate files. In respect of data pre-processing, all the normalization methods suited for gene comparisons within a sample are well tolerated by the algorithm. In the following table are listed the most common one:

Table 1 Common normalization method to account for gene comparisons within a sample

Normalization method	Description
CPM	Counts scaled by total number of reads
FPKM	Fragments per kilo base per million mapped reads
RPKM	Reads per kilo base per million mapped reads
TPM	Counts per kilo base of transcript per million mapped reads

A specific drop-down menu allows the user to select the tumour type corresponding to the expression profiles provided in input. This argument will instruct the algorithm on which model to use; if no action is taken, the algorithm will import the solid model by default.

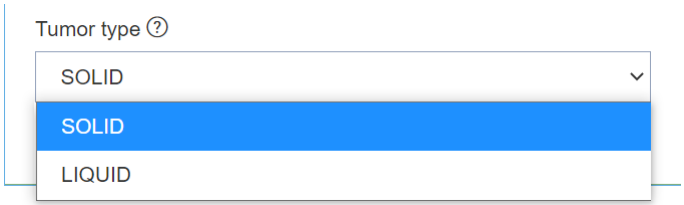

Figure 5: DpFrEP drop-down menu to select the tumour type.

Each of the solid and liquid models comprises two sets of drug profiles derived from the Broad and Sanger data releases. It is possible to selectively choose one of these sets of drug profiles from the specific drop-down menu.

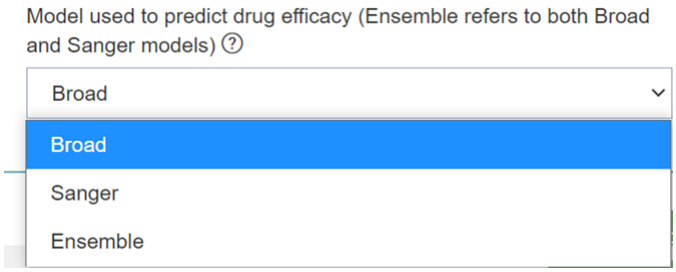

Figure 6: DpFrEP drop-down menu to select the model for drug prediction.

This argument is essentially introduced to reduce the time for computational analysis. It implies a knowledge of the drugs covered by the two Institutes and the expected performance on the paediatric cancer(s) of interest. To make the selection process easier, a metadata file containing the drugs covered by the two Institutes has been stored at https://github.com/inab/vre_dpfrep_executor/blob/master/dpfrep/data. In addition, PPV curves for six distinct paediatric tumours, including leukaemia (both acute lymphocytic and acute myeloid), Ewing's sarcoma, hepatocellular carcinoma, medulloblastoma and neuroblastoma, have been generated by alternatively employing one or the other model on the expression profiles of the corresponding CCLs. For those cancer types where one drug set does not outperform the other, or no benchmark has been carried out, it is recommended to select the Ensemble model that comprises both the Broad and Sanger models.

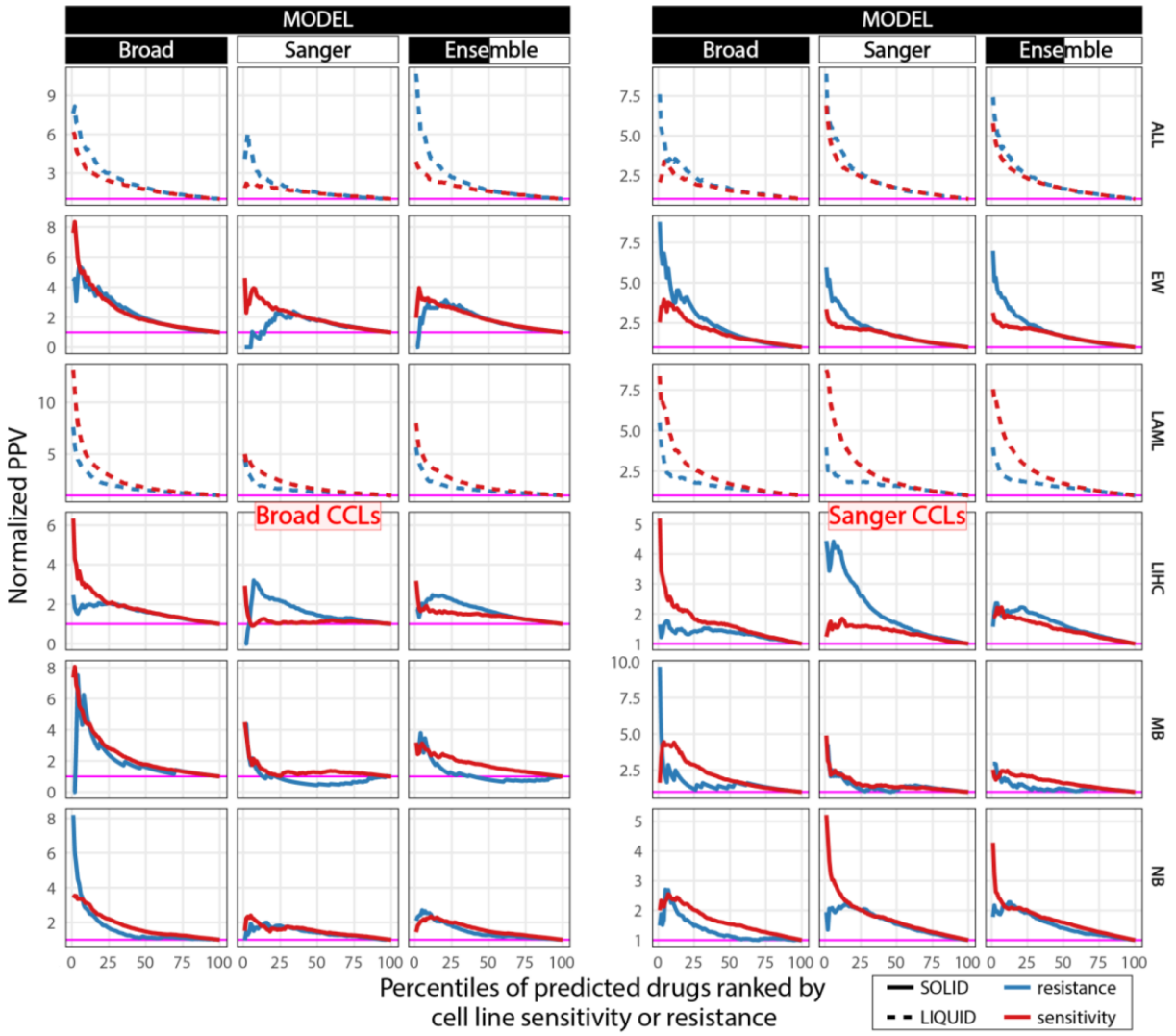

Figure 7: PPV curves for six distinct paediatric tumours. In the order acute lymphocytic leukaemia (ALL), acute myeloid leukaemia (LAML), Ewing's sarcoma (EW), hepatocellular carcinoma (LIHC), medulloblastoma (MB) and neuroblastoma (NB).

The software generates an XLS file in output. Sheets within the workbook are named using the sample identifiers provided in the input file. Each sheet contains the predicted drug ranked lists of the corresponding sample id, which includes the drug identifier, the drug name, eventually the drug target(s) and the sourced institute.

Chapter 4 Summary and Future Plans

We fully accomplished the objective of D8.1 to develop a software for predicting a list of small molecules with potential therapeutic application in childhood cancer. As a first benchmark, we evaluated the performance of our method using two publicly available datasets released by the Broad and Sanger Institutes (see Chapter 2). We are further validating the approach using specific paediatric cancer datasets, namely the TARGET (https://ocg.cancer.gov/programs/target) and the PPTC [6] datasets. In addition, a case study aiming to identify and prioritize drugs for hepatoblastoma (HB) patients is being conducted in collaboration with IGTP, which provided the expression profiles of these patients, and LMU partners. In particular, we applied our method to each patient, and we estimated similarities among the resulting drug ranked lists. Results, shown in Figure 8, highlighted that patients belonging to the same tumour subclass (either C1 or C2) are more similar, in terms of predicted drugs, than patients belonging to different subclasses. To identify drug-specific clinical pattern for each subtype, drug ranked lists from either C1 or C2 patients were aggregated using the Borda method [7]. Top predicted drugs for the C2 subclass (being the one with a more advanced tumour stage and the worst overall survival rate), are currently being tested in HB cell lines and PDX model. In future, we would like to upgrade the pipeline, developed for bulk datasets, to single cell data; this transition will lead to the development of approaches for compound combination modelling.

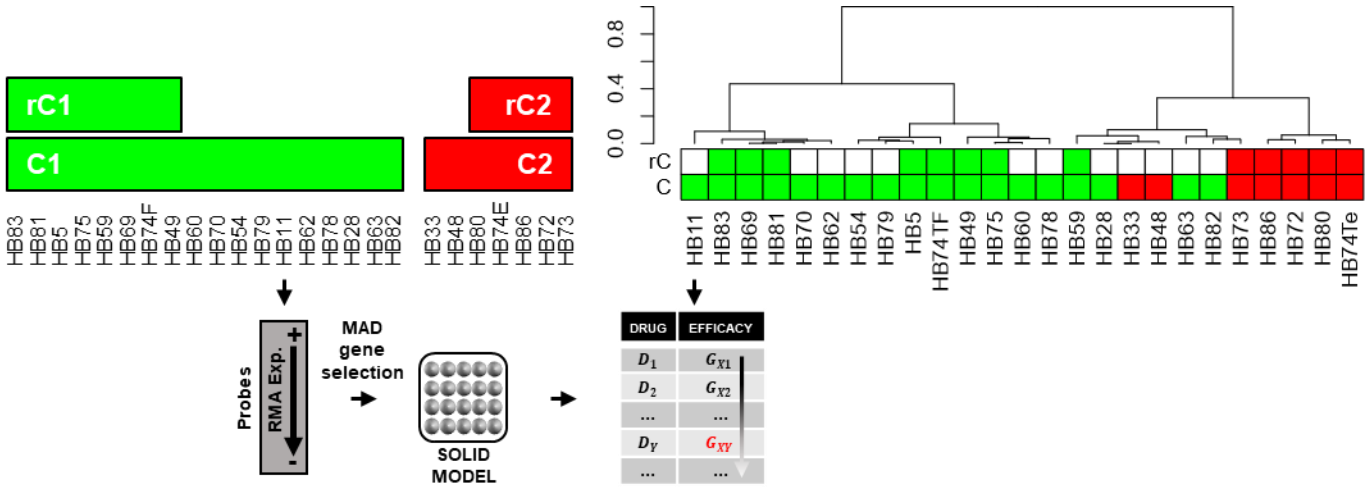

Figure 8: Clustering of HB patients. Expression-based (left) vs drug-based (right) clustering.

Chapter 5 Abbreviations

Abbreviation	Translation
AUC	Area under the curve
CCL	Cancer cell line
CSV	Comma separated values
DpFrEP	Drug prediction from expression profile
ES	Enrichment score
GSEA	Gene set enrichment analysis
HB	Hepatoblastoma
IC50	Half maximal inhibitory concentration
IGTP	Institut Germans Trias i Pujol
iPCCP	iPC central platform
LMU	Ludwig Maximilian University of Munich
PDX	Patient derived xenograft
PPTC	Pediatric preclinical testing consortium
PPV	Positive predicted values
TARGET	Tumor alterations relevant for genomics-driven therapy

Chapter 6 Bibliography

[1] A. Subramanian et *al.*, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–50 (2005).

[2] F. Iorio et *al.*, A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166**, 740-754 (2016).

[3] J. Barretina et *al.*, The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603–607 (2012).

[4] J. G. Tate et *al.*, COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Research* **47**, D941–D947 (2019).

[5] M. G. Rees et *al.*, Correlating chemical sensitivity and basal gene expression reveals mechanism of action. *Nat Chem Biol.* **12**, 109–116 (2016).

[6] J. L. Rokita et *al.*, Genomic profiling of childhood tumor patient-derived xenograft models to enable rational clinical trial design. *Cell Rep*. **29**, 1675–1689(2019).

[7] D. G. Saari, Explaining All Three-Alternative Voting Outcomes. *J. Econ. Theory* **87**, 313–355 (1999).