- Supplement S3 -
MERIDA: Methodological comparison to LOBICO

1 Data preparation

The idea of MERIDA (MEthod for Rule Identification with multi-omics DAta) is based on the LOBICO method by Knijnenburg et al. (Knijnenburg *et al.*, 2016). In this supplement, we provide a comparison of both approaches with regard to data preparation and ILP formulation. In particular, we highlight methodological commonalities and differences between the two methods. A general overview of the two methods is given in Figure 1.

Both MERIDA and LOBICO rely on binarized input and output data to generate logic formulas that can be used to predict the sensitivity of a cell line to a drug. The input for both methods can consist of arbitrary binary cell line features, e.g. one feature can represent the presence/absence of a specific mutation or the up-/down-regulation of expression in one gene. The output specifies for each cell line whether it is regarded as sensitive (1) or resistant (0) to a certain compound. However, the sensitivity data in the GDSC is given as logarithmized IC50 values, which are inherently continuous. Therefore, some thresholding procedure is needed to obtain binary values. To guarantee comparability of our approach to the approach by Knijnenburg et al., we generate these thresholds by applying the same procedure as described by Knijnenburg et al. to binarize the drug response data. This procedure results in an IC50 threshold for each drug that divides the cell lines into sensitive and resistant ones. MERIDA and LOBICO also use this threshold in their respective objective functions to obtain cell line specific weights that depend on the distance of IC50 values from the threshold. In general, it would also be interesting to use information from databases that store clinically relevant concentration values to define such thresholds, e.g. as given in Liston and Davis, 2017.

Generally, any logical model similar to LOBICO or MERIDA can enable the inclusion of a priori knowledge. However, to the best of our knowledge, this is not yet implemented for LOBICO. In order to enable the inclusion of a priori knowledge into the MERIDA model, we leveraged information from various well-known cancer-related databases: COSMIC (Tate *et al.*, 2019), CIViC (Griffith *et al.*, 2017), OncoKB (Chakravarty *et al.*, 2017), and the Cancer Genome Interpreter (CGI) (Tamborero *et al.*, 2018). Using these databases,

the contained biomarkers can be utilized to annotate if a particular alteration present in a cell line has previously been shown to be predictive of a positive or negative drug response. Furthermore, we used these databases to extensively annotate the oncogenicity status (e.g. *gain / loss of function*) for mutation-based input features. This additional knowledge facilitates the interpretation of the logic rules derived by our model. Details on the inclusion of a priori knowledge into the model input are discussed in the main manuscript as well as in Supplement S1.

2 ILP formulation

Figure 1 depicts a comparison of the ILP formulations for MERIDA and LOBICO. Both methods rely on Boolean logic to obtain a rule that describes the sensitivity to a particular drug. It can be seen that this rule is derived by combining features through logical AND and OR gates.

As stated in the objective function, LOBICO and MERIDA are formulated as a minimization problem that punishes false positive predictions (first term) and rewards true positives (second term).

The LOBICO formulation yields logic rules in *disjunctive normal form* (DNF), which allows the synthesis of any Boolean function from the ILP. A logical formula in DNF is a disjunction (i.e. logical OR connection) between conjunctions (i.e. logical AND connections) of literals. In LOBICO, K terms, t_1, ..., t_K are generated through conjunction of the input feature vectors or their negation (s_p and s'_p). The resulting K terms are then combined through disjunction to obtain the final prediction vector y'. Consequently, LOBICO relies on two hyperparameters: K, which is the number of terms, and M, which represents the number of elements (literals) per term. These parameters have to be determined in a tuning procedure such as cross validation.

While the LOBICO formulation does not restrict the resulting logical rules to a specific type of Boolean function, fitting such a flexible model is costly and may become computationally infeasible as K and M increase. For most of the drugs investigated in this publication we were not able to fit a LOBICO model with $M = 2$ and $K = 2$ in a reasonable amount of time (see Supplement S2). We terminated the training process, if the duration of fitting a single model in the cross validation exceeded six hours.

In MERIDA, the logical formula assumes a more restricted form consisting of only two terms, one term (s) that represents the collection of sensitivity-associated features (a_i) and the other one (r) representing the collection of resistance-associated features (b_i). Within each term, input features are connected through disjunction. The final prediction is obtained through conjunction of the sensitivity-associated term with the negation of the resistance-associated term and results in the prediction vector y'. The assumption behind this architecture is that a cell line is only sensitive to a drug if a sensitivity-associated but no resistance-associated alteration is present. As the shape of the logical rule is predefined in MERIDA, the space of possible solutions for the ILP is restricted in comparison to LOBICO such that more features can be included in the input and larger models can be build. In addition, we use only one hyperparameter M, which

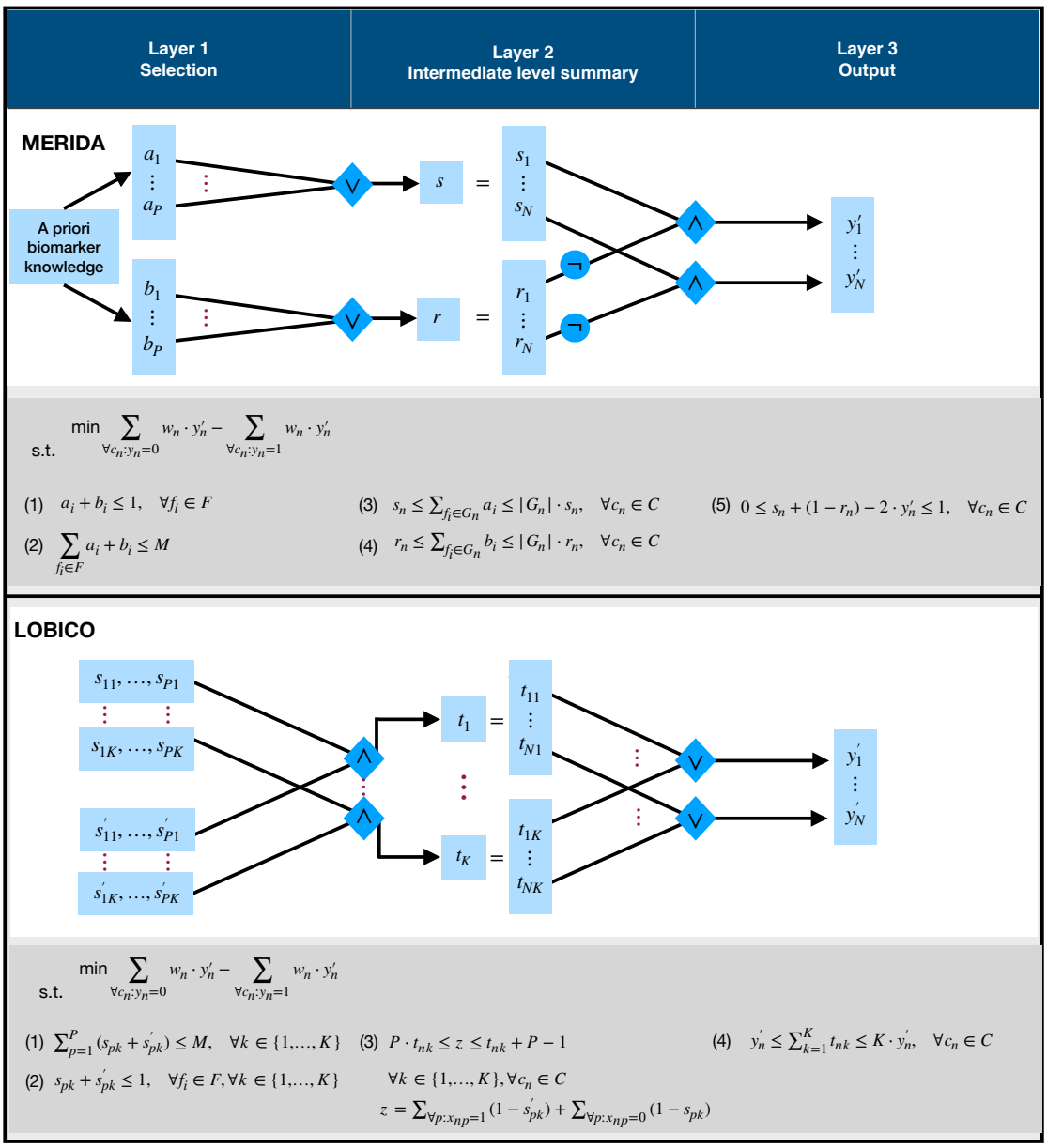

Figure 1: This figure depicts a summary of the ILP formulations for MERIDA and LOBICO including objective function and constraints. Here, $F = \{f_1, \dots, f_P\}$ represents the set of features and $C = \{c_1, \dots, c_N\}$ the set of cell lines. Both methods combine their respective selection variables via logical AND and OR gates to the final prediction.

denotes the total number of features to be selected for the logic rule. Consequently, the computational cost for hyperparameter tuning, e.g. through cross validation, is reduced compared to LOBICO.

Another factor that impacts runtime is the function that is used to weight cell lines in the objective function. While cell lines in the original LOBICO ILP are weighted based on the absolute distance of IC50 values from the binarization threshold, MERIDA also considers quadratic and cubic absolute distances. In our analyses, the quadratic and cubic weight functions improved the runtime of both MERIDA and LOBICO considerably, while selecting very similar features (cf. main text and Supplement S2). Adapting the weight function additionally allows putting more emphasis on cell lines that are further away from the threshold in model training.

An extension of MERIDA over LOBICO is the inclusion of a priori knowledge in form of known sensitivity-/resistance-related biomarkers into the model. We tested this feature extensively using various well-established cancer-related databases (CIViC, OncoKB, CGI, and COSMIC). While a priori knowledge can also be included into LOBICO, it is currently not implemented. Compared to MERIDA, integrating features as sensitivity-/resistance-related can be less practicable for LOBICO, as its ILP formulation comprises multiple terms in which a priori knowledge could be integrated, while MERIDA only uses two terms with predefined meaning. As a consequence, models with large values of K and M are probably needed for drugs with a lot of predictive biomarkers, which is currently prohibited by the high runtime of LOBICO.

In MERIDA, prior knowledge is added to the input matrix as additional features (for a detailed description see Supplement S1). The ILP then sets the specific variable for that feature to the predefined value, e.g. if a feature f_i is predictive of drug response, the corresponding variable a_i is set to 1. Note that a priori knowledge can not only be obtained from databases but also through iterative application of MERIDA, i.e. features detected in previous runs can be integrated as prior knowledge into the next run. This iterative application is permitted through the reduced runtime of MERIDA compared to LOBICO. Additionally, this iterative process can ensure that strong predictors stay part of the model.

Lastly, Knijnenburg et al. describe how LOBICO is able to ensure a minimum prediction sensitivity and specificity by adding two additional constraints to the ILP formulation. These constraints can also be easily integrated into our ILP formulation. However, we did not investigate the effects of such minimum requirement thresholds on prediction. Instead of adding constraints to the ILP for emphasizing sensitivity or specificity, the objective function can be modified, e.g. by incorporating a factor $\alpha \in [0, 1]$

$$\begin{aligned} \max \quad & \alpha \cdot \Big(\sum_{\forall c_n : y_n = 1} w_n \cdot y'_n - \sum_{\forall c_n : y_n = 1} w_n \cdot (1 - y'_n) \Big) \\ & + \\ & (1 - \alpha) \cdot \Big(\sum_{\forall c_n : y_n = 0} w_n \cdot (1 - y'_n) - \sum_{\forall c_n : y_n = 0} w_n \cdot y'_n \Big) \end{aligned} \tag{1}$$

However, this would introduce an additional tuning parameter.

References

Knijnenburg, T., Klau, G., Iorio, F. et al. (2016) Logic models to predict continuous outputs based on binary inputs with an application to personalized cancer therapy. *Sci Rep*, **6**, 1-14. https://doi.org/10.1038/srep36812

Liston, D. R., and Davis, M. (2017). Clinically relevant concentrations of anticancer drugs: a guide for nonclinical studies. *Clinical Cancer Research*, **23(14)**, 3489-3498. https://doi.org/10.1158/1078-0432.CCR-16-3083

Chakravarty, D., Gao, J., Phillips, S. et al. (2017) OncoKB: a precision oncology knowledge base. *JCO Precis Onco*, **1**, 1-16. https://doi.org/10.1200/PO.17.00011

Griffith, M., Spies, N., Krysiak, K. et al. (2017) CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Genet*, **4**9, 170–174. https://doi.org/10.1038/ng.3774

Tate, J. G., Bamford, S., Jubb, H. C. et al. (2019) COSMIC: the catalogue of somatic mutations in cancer. *Nucleic Acids Res*, **4**7, D941-D947. https://doi.org/10.1093/nar/gky1015

Tamborero, D., Rubio-Perez, C., Deu-Pons, J. et al. (2018) Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med*, **1**0 (25). https://doi.org/10.1186/s13073-018-0531-8